International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 10, Issue 5 (May 2014), PP.32-40

A Novel Approach for Improving the Recommendation System by Knowledge of Semantic Web in Web Usage Mining

[1]Nirali N. Madhak, [2] Chintan R. Varnagar, [3]Shahida G. Chauhan
[1] *Post Graduate Student,* [2, 3]*Assistant Professor,*
[1, 2, 3] *Department of Computer Engineering,*
Atmiya Institute of Technology and Science, Rajkot, Gujarat, India

Abstract:- The World Wide, We has influenced a lot to both users as well as the web site owners. Massive growth of World Wide Web increases the complexity for users to browse effectively. For increasing web site uses and to achieve desired goal efficiently, web server activities are hypothetical to be changed as per users' interests. To achieve this they have to record and analyze user access pattern which are captured in the form of log files. *Web usage mining* refers to process of analyzing interaction of user with different web application and deriving some important knowledge out of it. Web usage mining process produces result based on maximum usage history, stored in the web server access logs. So we propose the system, which uses semantic knowledge, derived from each page, along with knowledge derived from WUM. The system generates the most efficient and applicable recommendation list, and also provides equal opportunity for the content, which is added recently, to be incorporated in the list, if this best matches with users interest.

Keywords:- Data mining, Web usage mining (WUM), Web log mining, Content Recommendation, Recommender System, Web server log.

I. INTRODUCTION

The richness of information on the World Wide Web has attracted users to seek and retrieve information from the World Wide Web (WWW). When user is trying to access the content on web site, he/she is facing difficulty in finding fruitful content, that utmost match with the user interest. Recommendation systems are intelligent system that suggests and assist in selecting right content heuristically. Recommendation system is one of the applications of Web Usage Mining (WUM).

The main goal of the recommendation system in commercial domain is to improve website usability and thereby increasing the profitability of website owner. The patterns/ knowledge discovered provided as an input to the recommendation system, which recommends appropriate pages relevant to user interest.

Web access log file that resides in the web server, notes the client activity – request for accessing file on server initiated by client through the web browser. Web usage mining refers to the process of extracting knowledge/patterns by applying various data mining techniques like association rule mining (ARM), clustering, classification etc on the web access log files.

Recommendation systems, that makes suggestion based on only previous web access history or one that considers the similarity between item currently being accessed and items similar to that, fights with its own perils, and hence the recommendation generated does not guaranteed to be qualitative. The content of page i.e. overall theme, keywords, and its density should be considered in deciding the candidature of the content, to be presented in the recommendation list.

Here we propose content recommendation system which gives, suggestion based on not only access history of current and other users, but also considers the semantic knowledge achieved, from web content mining. The architecture presented, proved, to be giving better and optimized results. Additionally the system is prone with the addition of newly added page or unvisited pages which matches with user interest.

So that proposed solution is combination of two web mining approaches that are Web Usage Mining and Web Content Mining (WCM), **Section IV**.

Section II explains the brief about work done so far. **Section III** explains the input for web usage mining and on various **Pattern discovery** techniques that can be applied on preprocessed web access logs gathered to mine knowledge from it. **Section IV** discusses on proposed content matching algorithm which uses the heuristic function to build the recommendation list, **Section V** concludes with the merits of proposed approach and future scope.

II. LITERATURE SURVEY

Data mining is the process of extracting previously unknown information from different types of data like text, audio, video etc., which leads to fruitful knowledge. Now a day's web has proved to be as affluent sources of data where multiple domains are accessed and mined, mining web data is referred as **Web Mining**.

Web Mining can be broadly divided into three phases: [7] Web Usage Mining (WUM), Web Content Mining (WCM) and Web Structure Mining (WSM). In this paper, we propose a recommendation system, which not only considers user navigation history but it also considers content of web page. So, here proposed approach is a combination of WUM and WCM, which in combination can give better recommendations.

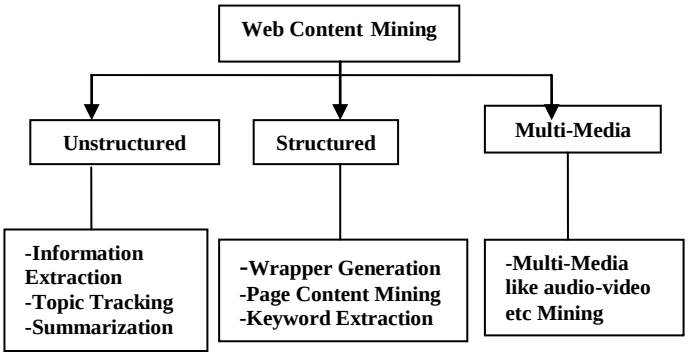


Fig. 1: Classification of WCM

As shown in figure 2.1, WCM aims to extract information from different type of data of web page. By applying various techniques of web content mining on unstructured data ,a HTML pages, is challenging task because, HTML web pages have multiple tags which make the web pages highly unstructured. Topic tracking is the technique by which a registered user can track the topic according to his/her interest. He/she have to register with the topic, whenever there is an update or news regarding the interest of the user happens, he/she will intimate by message.

Structured data extraction is the process of extracting information from the web pages; programs are written which helps to extract such information, which is surrounded by Wrapper. By the keyword extraction technique which is applied on the structured data gives the information about the number of keywords which is of one phrases, two phrases etc.

Multimedia data extraction is the process of finding interesting knowledge from data like audio, video, image and text. By developing methods and tools to organize, supervise, search and to perform domain specific tasks for data of different domains such as surveillance, meetings, broadcast news, sports, archives, movies, medical data, personal and online media collections

There are various approaches used for web content mining like, by using the ontology's which specify the domain specific knowledge and finding its relevant content, by applying the various clustering techniques. Ontology is a domain specific knowledge that could be used to describe information on the Web. Ontologies based web mining can be used to improve search to web data by adding Ontology explanation, better browsing capabilities and personalization of Web data from the user's interest profile. With the use of various domain specific knowledge (ontology) ,we can improve the user interest profile integrating with the semantic web approach.

Semantic web is all about integrating and extracting intelligent information from structured or unstructured data from the web and the knowledge which is embedded in web applications, thus providing a semantic-based access to the Internet. Ontology is one of the layers of Semantic Web's architecture as proposed by Sir Tim Berner's Lee.

Accordingly [], Jayatilaka A.D.S!, and Wimalarathne, It provide idea how ontologies are used in semantic web. Semantic web mining is relatively new sub-field of data mining. It has a vast scope for investigation keeping in view of the availability of tons of unstructured data on WWW.A user-oriented semantic web search is the need of today and days to come. This field if explored in a right manner will provide unlimited opportunities to extract knowledge to ultimately improve the profitability of an individual or company by mining knowledge from unstructured and/or structured data available across the internet.

Jayntilaka et al. [14], insisted the concept of semantic web using both web author's view point and web user's perspectives in the ontology learning process. It also eliminates the use of web site dependent characteristics in extracting semantics.

The extracted concepts with help of ontology web language and the conceptual relationships gives rise to a semantic network and form ontology. There are three main stages in this method which are:

i) Concept and conceptual relationship extraction through web content mining,
ii) Conceptual relationships identification through web usage mining
iii) Refining/Merging the conceptual relationships obtained through the web content mining process

[15]Web mining based on semantic networks is used the new semantic to improve the Web mining result. With the analysis of semantic level on RDFMS resource description, RDFMS hierarchical clustering method based on semantic distance data is proposed. The inductive logic programming design is proposed for the semantic Web data mining technique and how to apply this technological algorithm description.

Jenice Aroma R.et al.[16], proposed the semantic discovery algorithm combines the method of semantic similarity measure between words to be applied with the queries supplied, in order to retrieve the semantically matched results. To optimize the retrieved results, ranking be applied over the matched results. It brings more relevant results to be ranked highest. Thus, Intelligence on Information retrieval for achieving more relevant results can be implemented on applying this proposed semantic discovery algorithm with semantic similarity measure. But for that we have to classify each document into its appropriate ontology classes.

Content Recommendation Systems are one of the applications of WUM as well as WCM. Content recommender system, gives suggestion based on access history of current user and others.

The main objectives of the system is to improve usability of web site by dynamically and automatically understanding and modeling visitors navigational behavior to build user profile ,***second*** exploiting thus created knowledge base by the application of most suited heuristic techniques, often embedded within a special component called intelligent agent. There by recommending appropriate pages and gaining user satisfaction and easy of surfing [9]. Identifying and selecting proper recommendation to user is complex process and mostly the techniques applied are heuristic in nature.

A Recommender System, merely applying various data mining techniques to the web server log data fight with its own perils, as it will not help to derive complete, accurate and efficient recommendations. First Information in the web log is very limited, Second it assumes that the requests are fulfilled sequentially, which is not true for con-current information need, it means user can fire the requests parallel. Finally if user leaves after visiting few links, without competing transactions, it might mislead our results. But visitor may have left because of non fulfillment of his information need, poor link navigation or bad navigation also [17].

Ting CHEN et al. [18] suggested a system, which does recommendation, consisting of three tiers (layer). L-1(Layer-1) is row information collection agent, which collects data from client machine. L-2, a logic layer uses this data to create Dynamic User Profile (DUP), L-3 is responsible for presentation and customized UI. [18] Suggested to build such a dynamic profile from various hardware level events like keyboard, mouse etc.

According to [19] to discover sequential access patterns first association rule mining algorithm (A-priori) and its associated modified algorithm, second methods based on stochastic probability, Markov chains, third to use weighted association rule mining, which allows different weights to be assigned to different items, hence improving AR model.

[4] Suggested similar concept, which creates user groups (clusters) and uses intelligent recommendation agent, covering all type of multimedia contents. This is perfect combination for the integration of user specific activity (service) recommendation in social networking and its associated web services, in the environment of web 2.0.By the above literature survey we will finding some of techniques for web usage mining and web content mining. It is possible to optimize the recommendation system by discovery of new algorithm and new approach to measure the result of combination of both technique WUM and WCM as discussed in Section IV.

III. INPUT FOR WEB USAGE MINING

In the last decade, many researchers have developed many different kinds of approaches of web system to achieve web personalization. While surfing the web sites, users interaction with web sites are recorded in web server access log file.

There are three main sources to get the row log data [1], which are namely:

1) **Web server log file** data like Access Log, Agent Log, Error Log, Referrer Log.
2) **Client side log files**, which are more authenticate and accurate.
3) **Proxy server or firewall log files,** which contains access log captured at organizational gateway, may be varied in format, content and in other aspect from server to server or across different software and hence very difficult to get useful information from it.

Web Server access log data:

The most frequently used source for web usage mining is web server access log data. This web log data is generated automatically within web server when it services any request sent by the user, which contains all information about visitor's activity.

Accordingly Suneetha K R et al.[1] proposed various preprocessing techniques apply on web access log.

Many different formats for web access log data are available like:-

1. Common log format
2. Extended common log format,
3. Centralized log format
4. NCSA common log format
5. ODBC logging

Among all common or extended file format is mainly implemented by web server due to many reasons.

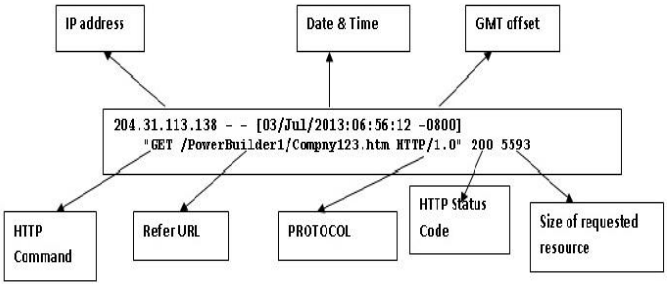


Fig. 2: Attributes of ECFL [7].

Extended Common Log File Format (ECFL), Figure 2.2 is important in web usage mining, as it can be customized as per the requirements and is followed by most of the web server. The additional attributes that are captured are

i) REFERER_URL define the URL where visitor came from
ii) HTTP_Command reflects GET/POST method
iii) PROTO--type of protocol used for the request
iv) HTTP_Status-status code generated by the request
v) GMToffset-signed offset from Greenwich Mean Time

[1] Web access logs may be used to increase the effectiveness of web portals or for better understanding of user behavior. The main research challenges of this field are identifying the non human entries made by web robots, design efficient heuristics for user session identification and finding association among different user's access patterns.

The system takes raw log data for the website http://eyuva.com for the periods between 2014-03-20 [05:35:14] TO 2014-04-23 [4:41:59].As shown below:

Client side log data*:*

It is refer to recording of activities, events that happens within the premises of client machine. Like mouse wheel rotation, scrolling within a particular page, mouse clicks, content selection. In some case it is advantageous, as it eliminates necessity of session identification, caching [11]. This can be recorded by number of ways:

1) **By integrating java applet with web site**: Java applet records each of the activity of users. But for that java plug in need to be installed on each client side browser. Also user may experience delay in page loading time, when applet is loaded for the first time [11]. Additionally all current pages of the website need to be redesigned and recreated in terms of Applet.

2) **By writing Java Scripts**: Java Script needs to be inserted at appropriate places, need to be invoked as and when required (need to be associated with appropriate event handler), in each page of web site. This will record these interactions of user with web page and report it to server when transaction is complete.

3) **By developing a browser plug-in(extension):** Which need to be installed only once which can record this kind of interaction and will send the record at finite interval of time or just before when user is about to close the connection with website or when user is quitting from browser. This can be done without changing the underlying design, architecture or technology of web site.

Proxy Server Log Data:

At many places network traffic is routed through a dedicated machine known as a proxy server, all the request and response are serviced through this proxy server. Study of this proxy server log files, whose format is same as of web log file may reveal the actual HTTP requests coming from multiple clients to multiple web

servers and characterizes, reveals the browsing behavior for a group of anonymous users sharing a common proxy server [11].

Some web sites use n-tier architecture to have reliable, efficient and secure web applications. Log data that are gathered at application server while servicing the users request can also be used for web usage mining. They peculiarly show how user requests are serviced and may assist in identifying and understanding the internal calls-page access resulted to fulfill a single request.

IV. PROPOSED CONTENT RECOMMENDATION SYSTEM

In the literature survey till now I have not came across the method, technique, or approach which uses the content of the page ,as a one of the parameter to decide whether it should be made available – get listed in the recommendation list or not. So by considering this parameter in building recommendation list will not only improve the quality of recommendations made by system for user – depending on his till now browsing patterns (interest) but will provide a equal chance for the pages (content), which are added recently in the system.

At no place up to authors knowledge it is think of, no argument is given in support or favor for using both ,i.e. web server access log and content of each page to build the recommendation set. Hence the system that knows the contents of the pages that user has browsed till time, will surely be able to predict the most interesting page – content, that user is about to like or will be interested in or he is looking for.

So content can be more precisely be predicated by modifying heuristic algorithm by considering this semantic knowledge – that reside within a page , when combined with other traditional methods of content recommendation like frequent access pattern mining using Apriori algorithm or some other techniques to mine a association rule and combine that all results according to heuristic function.

Proposed system architecture for content recommendation system can be logically divided in to TWO co-related PHASES named. Each phase is logically co-related with each other in a sense it accepts output of other as an input. But from the viewpoint of their functions and their relative execution domain it can be classified in above two phases.

1) Back End (Offline) and
2) Front End (Online).

Phase-I (Back End):

This Module is responsible for capturing and storing the content of the pages, in a meaningful and easily accessible – retrievable way known as web content mining. I.e. getting to know the semantic of each page and to store it in a way, that is most easy to access as and when it is required.

It also processes web server access logs to mine frequent access patterns, which in itself is not an easy task and requires a series of steps known as data preprocessing, pattern discovery and pattern analysis

Back end phase processes the data which are relatively static in nature like web pages – content of the web site and web server access logs. They are static in a way as it does not need to compute or have to perform any processing-each time user request is fired or recommendation is to be generated. However it need to be updated periodically, so as to get qualitative and most accurate, up to date results. So frequent run of this phase is necessary especially when new content is added in to the web site and also after finite time even though it is not added or updated so as to accommodate and mimic the correct current trend of the website user.

Back End Phase further can be classified based on the function and their operation domain, in following Modules.

A) Web usage mining on web server log
B) Web content mining

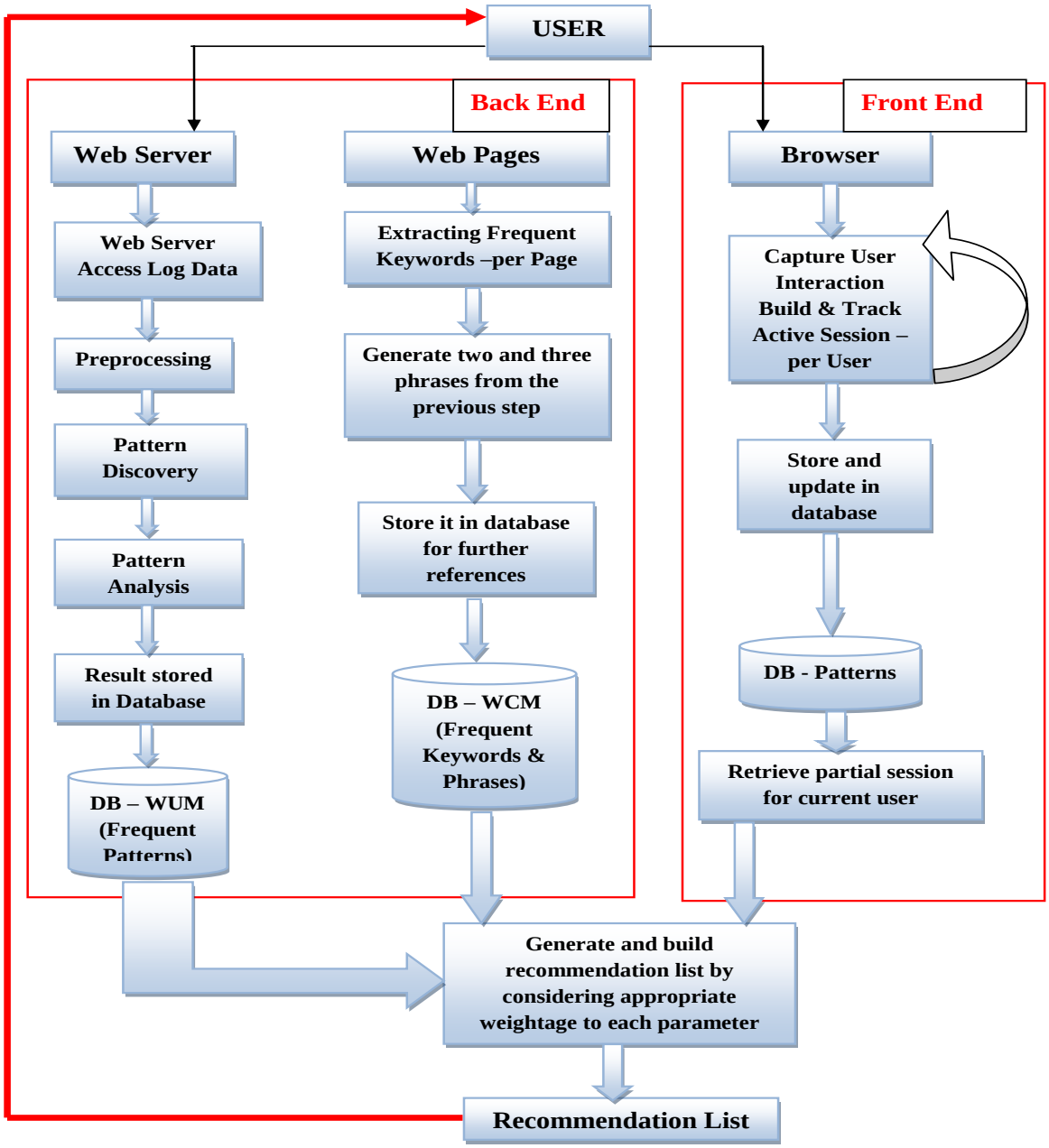


Fig. 3: Proposed Architecture of Recommendation System

A) Web usage mining on web server log:

This module uses web log files as input and derives some meaningful knowledge, frequent access patterns out of it. Entire process of web usage mining can be logically divided into four significant and co-related steps, which are Data Collection, Data Preprocessing, Discovery of Pattern, and Analysis of pattern. For that first web raw log data need to be cleaned, it means removal of unwanted rows, known as data preprocessing. Secondly to mine frequent access patterns, using technique of Association Rule Mining, specifically a simple Apriori algorithm is applied and results (patterns) are stored in some permanent storage for referring to it later.

Data Preprocessing: Data Preparation is the most complicated and time consuming task. About 80 percentages [2] of time is given on this process to strengthen quality of data because as qualitative the data is better the results. For this data preparation task which mainly includes various sub-task namely data cleaning, user identification, session identification, path completion and transaction identification [6].

Discovery of pattern: It is the ultimate stage where some useful knowledge will be derived by applying various statistical and/or data mining techniques at hand from various research areas like data mining, machine learning, statistical method and pattern recognition. Frequently used techniques are classification, clustering, association rule, sequential pattern etc [12].

Association Rules are able to discover related item occurring together in same transaction, and is used to find interdependency, co-relation among the pages. Such number of rules generated could be very large so two measures support and confidence is employed, which determines importance and quality of rules [1]. A-Priori and its many versions are developed to mine association rule.

B) Web Content Mining: This step is responsible for generating keywords – which are appearing most of the times from the contents of each web pages of a particular web site, and then it generates two word phrases and three word phrases by applying following proposed User_Interest_Content_Matching algorithm, which helps to optimize the results of recommendation list.

Algorithm**: User_Interest_Content_Matching**

Step-1: For each webpage extract top N keywords, whose frequency is maximum.
Step-2: Generate two words phrases and three words phrases from the keywords generated at step-1.
Step-3: Store these result in database for further reference.
Step-4: Apply the member function and compute the value for ***each web page*** from the set of recommendation list (web pages) as follows:

Fwi= ***A * no. of occurrences of single phrase keywords that is common across all the visited page till now***
+
B * no. of occurrences of two words phrase keywords that is common across all the visited page till now
+
C * no. of occurrences of three words phrase keywords that is common across all the visited page till now
+
X * support count of (% times) the page is referred in history
Where A, B, C Є [0, 1] & X Є [1, 100]

Step-5: Choose the best promising page (content) based on the combination of usage history obtained and appropriateness of content as follow:
→Compare the values computed by above heuristic function, computed for each page and choose the maximum value – which indicates most likeliness of the value to be presented i.e. which user might like the most.
Step-6: Display Recommendation set based on descending value of heuristic function.
Step-7: Exit

By applying the above algorithm that helps to find heuristic function, that will help to combine two different approaches web content mining and web usage mining.

Phase-II (Front End):

It is responsible for processing of a URL request of user by Personalized & Interest Specific Intelligent Recommendation Agent. These results (Recommendation Set) can be made available to the user, which is based on the current browsing history of current user that too without changing neither the page code (to insert the recommendation results) nor browsing experience of a user is affected for the website under consideration. This data will be supplied in the side bar, also known as a slider, without restructuring the web page. To build the final recommendation list algorithm Generate_ recommendation _list is applied.

Algorithm**: *Generate_ Recommendation List***

Step – 1: Retrieve IP address of client requesting for the resource.

Step – 2: Capture and build the active session i.e. list of web pages traversed so far, for all the currently active user dynamically, by applying session identification technique.

Step – 3: The new current, partial session will be compared with already existing aggregate usage profile from the frequent access pattern set stored in knowledge base.

Step – 4: Build a recommendation set based on ascending value returned by Heuristic function Fwi of procedure User_Interest_Content_Matching.

Step – 5: If more than one pattern found with same support value for current partial session, second measures like page weight, computed by considering various parameters, stored in database will be used to decide the priority for displaying the results.

If we are going to consider only the web usage mining behavior of user ,how many pages are frequently access by multiple user we are getting result by considering active time spent by user on the web page as below:

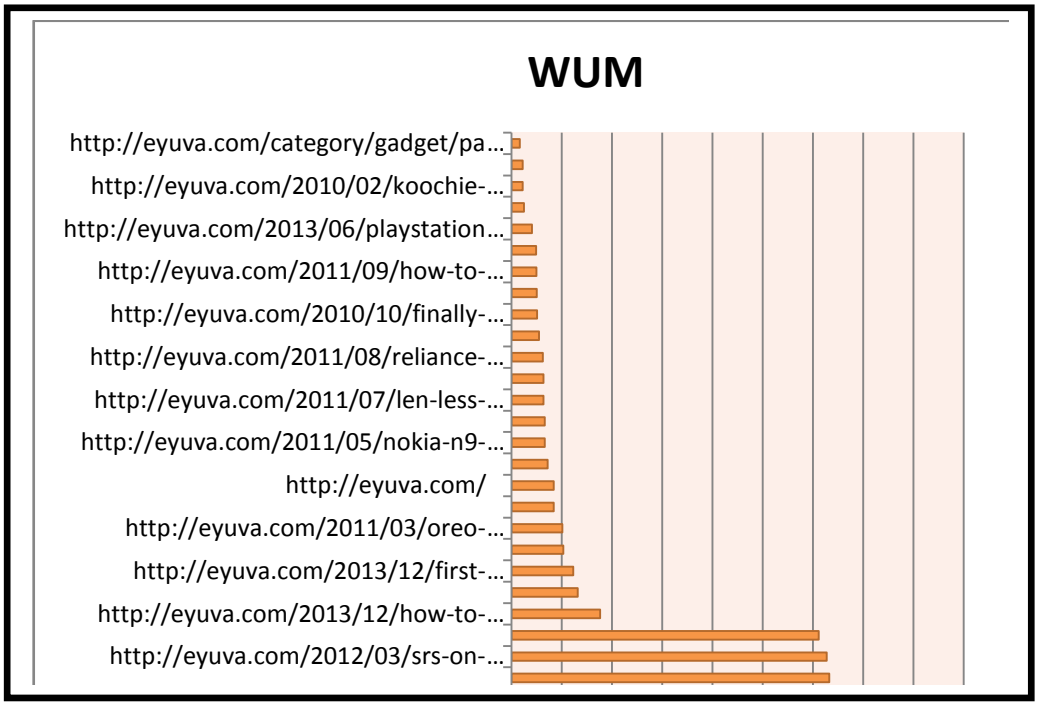
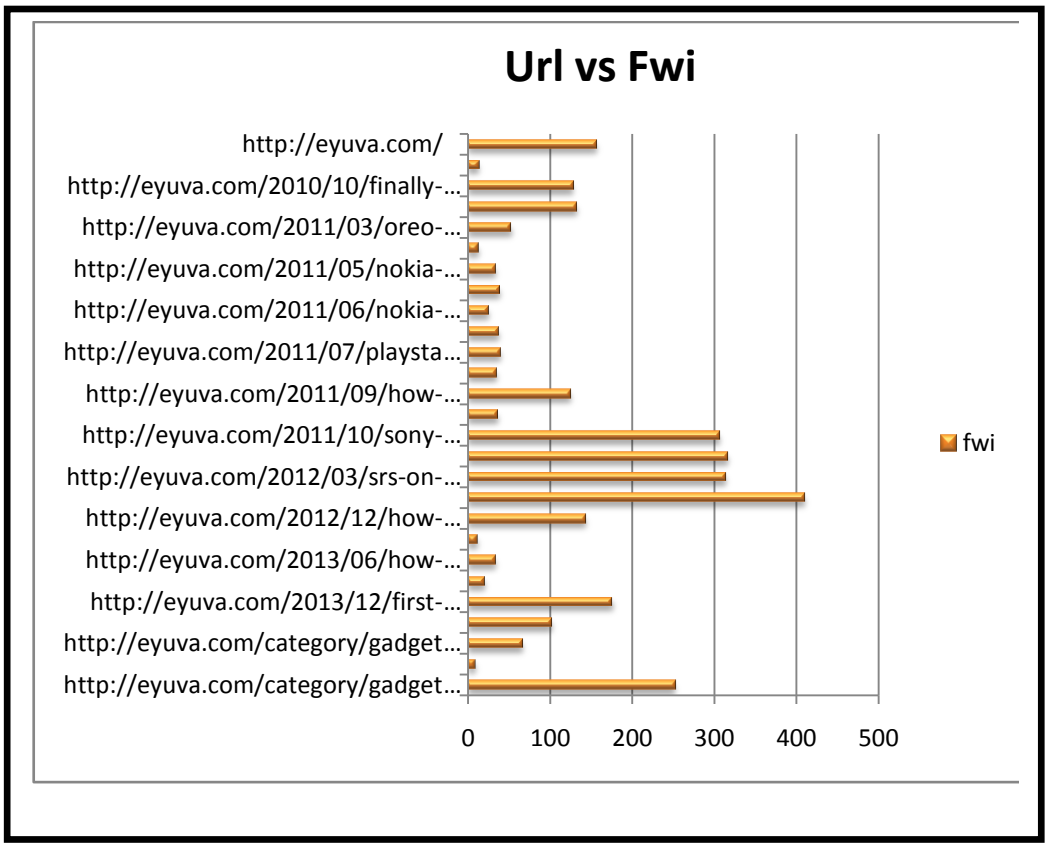


Fig.4: URL vs Accesstime

By applying the for generating recommendation list according the ascending value returned by heuristic function. Following figure shows the how to index url with the value of Fwi.



By comparing results of the only web usage behavior and combination of WCM and WUM, Second approach gives us fruitful results that are beneficial to finding user interest web pages that best matches with user interest profile.

V. CONCLUSION

Web sites are great amount of use for the user. Web sites are built, deployed and maintained to serve with various function to user. Web is increasing its importance in each possible aspect and is becoming an expected part of one's routine and regular resources. Hence there are sufficient opportunities and wide scope and requirement to study this field in the depth. Systems incorporating knowledge from ONLY navigational history (WUM) often produce incomplete, inefficient result – as it is based on single parameter .So newly added content(web page),will not be listed in Recommendation List , although it best matches with users interest – just because so far it is not visited or visited very less time. So a system can be improved, if it considers semantic knowledge of each page and incorporates this factor, with knowledge achieved from WUM, dynamically – for each possible element (page) from recommendation set.

REFERENCES

[1] Hauqiang zhou and Hongxia Gao et al. "Research on improving Methods of processing in Web log Mining", IEEE, 2010.

[2] R. Cooley, B. Mobasher, J. Srivastava, "Web mining: information and pattern discovery on World Wide web", tools with artificial intelligence, Ninth IEEE International November 1997.

[3] J. Srivasta, R.Cooley, M.Deshpande, P.Tan, "Web usage mining: discovery and applications of usage patterns from Web data", ACM SIGKDD Vol.7, No.2, Jan-2000.

[4] S. K. Pani et al.," Web Usage Mining: A survey on pattern extraction from web logs", International Journal of Instrumentation, Control &Automation, Vol.1, Issue 1, 2011.

[5] Ting Chen et al., "Content Recommendation System based on Private Dynamic User Profile", VIth International Conference on Machine Learning and Cybernetics, IEEE, August-2007.

[6] J. Srivasta, R.Cooley, M.Deshpande, P.Tan, "Web usage mining: discovery and applications of usage patterns from Web data", ACM SIGKDD Vol.7, No.2, Jan-2000.

[7] Chintan.R.Varnagar et al.,"Web Usage Mining:Asurvey on Pattern extraction using web logs",IEEE,2013

[8] Sang-il hwa-sung kim et al.,"Ontology Modeling for provision of semantic based open API", IEEE, 2013

[9] Rana Forsati et.al, "An Efficient Algorithm for Web Recommendation Systems", IEEE, 2009.

[10] Ravi Bhushan and Dr.Rajendra Nath,et al., "Automatic Recommedation of web pages for online users using Web Usage Mining", IEEE, 2012

[11] Saim Shin et al," The User-group based Recommendation for the Diverse Multimedia Contents in the Social Network Environments" , IEEE, 2011.

[12 Jenice Arona R,Mathew kurian,"A semantic Web:Intelligence in information Retrieval",IEEE-2013.

[13] Yanjing Zou,"Personalized Automatic Recommedation for the web based Autonomous language learning system based on Data Mining Technology",IEEE-2011.

[14] Jayatilaka A,D.S, "Knowledge Extraction for semantic web using Web Mining", IEEE, 2013.

[15] Zahid Ansari,A.Vinay Babu, "A Fuzzy set theoretic approach to Discover user Session from Web Navigational Data ", International Conference on Control and Automation, IEEE, 2011.

[16] Liu Kewen, "Analysis of Preprocessing methods for web usage mining", International Conference on measurement, Information and Control, IEEE, 2012