Web Personalization System: Hybrid Approach

Annapurna Pawar[1], Pankaj Waghralkar[2]
[1, 2] Department of ME(SE)
[1, 2] MIT, Aurangabad, Maharashtra, India

Abstract-*A web personalization system provides most relevant pages to user according to the user interest domain. The system obtains the knowledge user interest domain by analyzing the user browsing history. The browsing history is obtained by accessing the web log data at server and cookies. The web personalization system is also known as recommender system. The motivation for this hybrid approach comes from the observation that the existing system provides the relevant data, but the system should be effective and fast. To achieve this, it combines usage mining along with content mining. This approach improves the system performance with the help of content filtering. Thus user can obtain the relevant information on web as per user interest.*

Keywords-Frequently accessed patterens, DOM, Clustering, Page ranking.

I. INTRODUCTION

Personalization is the ability to provide the relevant information based on the user interest. The main goal of personalization is to help users find the information they are interested in. Most of the personalization systems try to filter available content found potentially interesting for that particular user. Extraction process information from the log files of website isused to identify the usage patterns and profile of user.

Web personalization system incorporates the web mining concepts: Web content mining, web structure mining and web usage mining. The motivation for this hybrid approach comes from the observation that personalized content on the web is relevant. Nowadays, when information overload is one of the common problems of web uses, it is difficult for the users to find the relevant information. This issue is solved with personalization. Personalization is used to provide the relevant information on the web. It provides information based on the user browsing history.

Knowledge obtained by studying the preferences of web users can be used to improve the effectiveness of the website. More web services are interested in learning user interest, so they can better target the product according to user interest. This paper is aimed to identify the frequent patterns from weblog data using Apriori algorithm. Patterns will be analyzed for information from data. Then by applying the content search by using wrapper generation and DOM construction it makes the cluster according to the content. Each cluster is get assigned with the cluster grade and each page in that cluster is get assigned with page rank based on the number of visting by the user for specific page. The following section discusses Methodology which can be used for the system development.

II. ARCHITECTURE

The web personalization system architecture consists of the 5sub-systemswhich are the parts of system, represented as modules. These sub-systems are: Sequential analysis, clustering, content filtering, Content Caching and page rank. An overview of the architecture of the proposed system is given in Fig.1.First, all users' web access activities of a website are recorded by the WWW server of the website and stored into the Web Server Logs. Each user access record contains the client IP address, request time, requested URL, user ID, HTTP status code, etc.

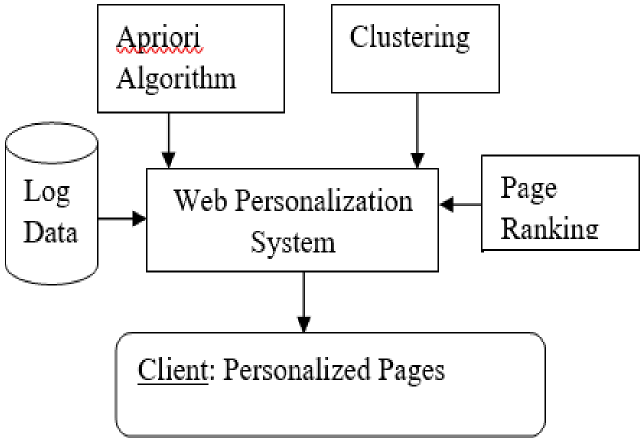

Fig.1 System Architecture of WPS

When a user visits the website, the user's HTTP requests in the current browsing session are recorded in order, and the current access sequence is constructed. Each user accessing the website can be identified using his/her IP address [1, 3]. The system accesses the data from the log data then it applies the sequential pattern analysis for finding the succeeded are the pages which are accesses sequentially. Clustering applies for making the classification of the web pages according to the content and page rank.The content filtering is used to filter the content which is approximate. The finally page rank and is used to number or weight page according to number of visitors and the content is stored to the cache memory.

III. METHODOLOGY

For this system we use input as the web log data and it analyses the data using sequential pattern analysis using Apriori algorithm. Sequential pattern analysis is: given a database of sequences where each sequence is an ordered list of user access based on access time and each access consist of a collection of useful information, and then searched the entire access pattern with minimum support by the user, where supports a number of database sequence contain the pattern. After sorting the pages it applies the content search by crawler. Then makes the cluster according to the contents. Each cluster is get assigned with the cluster grade and each page in that cluster is get assigned with pagerank based on the session time of the user for specific page. The pages after the sequential analysis, it crawls for the content with the help of crawler or by Wrapper generation. Web crawler takes data form user, search it and get well selected pages using breadth first search algorithm. In Wrapper generation, web page data can be extracted using HTML wrapper. Here the data is DOM tree which is constructed by web browser [1, 2, 6].

The training data is get find out from the extraction of content by crawling. Based on this content or pattern, a system does cluster the data by applying the association rule mining and by construction of pattern tree and by using the K-means algorithm (EM). This cluster is based on content that represents the interest of the user. Then each cluster is assigned with cluster grade for getting distinguished from other cluster. So each cluster represents the different unique interest domains. The pages in each cluster are assigned with numerical weight based on the number time for which the page is accessed. Then at final, pages with rank are get displayed to user as the personalization result based on the user interest based on user browsing history.

A. Apriori Algorithm for frequently accessed patterns:

```
Pass 1 Generate the candidate itemsets in $C_1$
          Save the frequent itemsets in $L_1$
   Pass k Generate the candidate itemsets in $C_k$ from the frequent
   itemsets in $L_{k-1}$
            Join $L_{k-1}p$ with $L_{k-1}q$, as follows:
            insert into$C_k$
            select$p.\text{item}_1, p.\text{item}_2 . . . p.\text{item}_{k-1}, q.\text{item}_{k-1}$
            from$L_{k-1}p, L_{k-1}q$
            where$p.\text{item}_1 = q.\text{item}_1, . . . p.\text{item}_{k-2} = q.\text{item}_{k-2}, p.\text{item}_{k-1} < q.\text{item}_{k-1}$
            Generate all ($k$-1)-subsets from the candidate itemsets in $C_k$
            Prune all candidate itemsets from $C_k$ where some ($k$-1)-subset of the candidate itemset is not in the frequent itemset $L_{k1}$
      Scan the transaction database to determine the support for each candidate itemset in $C_k$
      Save the frequent itemsets in $L_k$
```

Association rule describes analyzing and presenting strong rules discovered in databases using different measures of interestingness. The support supp(X) of an item set X is defined as the proportion of transactions in the data set which contain the itemset. In the example database, the itemset {milk, bread, butter} has a support of 1 / 5 = 0.2 since it occurs in 20% of all transactions (1 out of 5 transactions).

B. Clustering:

To extract information fromdeep web that is large collection of dynamic queryable databases, we need a system that can extract automatically. For this purpose ewe use web content mining techniques that uses XML version of HTML query interfaces. Web content mining is a form of text mining and can take advantage of the semi structured nature of web page text.

In Wrapper generation, web page data can be extracted using HTMLwrapper. Here the data is DOM tree which is constructed by web browser.DOM is a standard language that gets a web page as an input and shows it in a structured tree from interfaces, objects and relations between them as an output.

The training data is get find out from the extraction of content by crawling. Based on this content or pattern, a system does cluster the data by construction of pattern tree and by using the K-means algorithm. In statistics and data mining, k-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. This results into a partitioning of the data space into Voronoi cells.

((f1log(n/df1), (tf2, log(n/df2), ...,(tfn,log(nldfn)).

Where tfi is the frequency of the ith term in the document and dfi is the number of documents that contain the ith term. To account for documents of different lengths, the length of each document vector is normalized so that it is of unit length(||dtfidf=1||).The most common algorithm uses an iterative refinement technique. Due to its ubiquity it is often called the k-means algorithm; it is also referred to as Lloyd's algorithm particularly in the computer science community.

C. Ranking the Page in Cluster:

The pages in each cluster are assigned with numerical weight based on the number time for which the page is accessed. This system will refer the weighted page rank algorithm. The weighted Page Rank algorithm (WPR), an extension to the standard Page Rank algorithm, is introduced.WPR takes into account the importance of both the inlinks and the outlinks of the pages and distributes rank scores based on the popularity of the pages.

IV. PERFORMANCE EAVALUATION

The algorithms will conduct the Implementation experiment in NetBeans 8.1 IDE,with Java programming language is used. While the front end is in NetBeans framework the backend i.e. Database is in SQL Yog community. This application can be accessed by any web browser. The Results are successfully displayed on many web browsers like Internet Explorer, Mozilla Firefox, and Google Chrome.

The observation of the overall performance based on the Usage mining algorithm used which is Apriori algorithm and the content mining algorithm K-means. To evaluate performance first In this research evaluated performance for system only on usage mining algorithm which is Apriori, the result of this will be number of accessed links provided to user as for recommendation and then this result will compared with Hybrid System. Performance analysis will be conducted on real data sets. For now the impact of parameters to assess are i.e. Support, Clustering and Page rank.

The data in this research taken of the 20 users who accessed the system and visited the links provided. In this research provided the 6 links to the user. The data set contains the details of web log i.e. user name , Ip address , time , session, link visited etc.

Scenario 1:- By applying Apriori algorithm in this research have the sequences of the users visited pages along with the machine address, time of visit. In this scenario it gives the accessed link with as the recommendation. Here it gives maximum links to only 1 user and minimum links to all i.e 20 users.

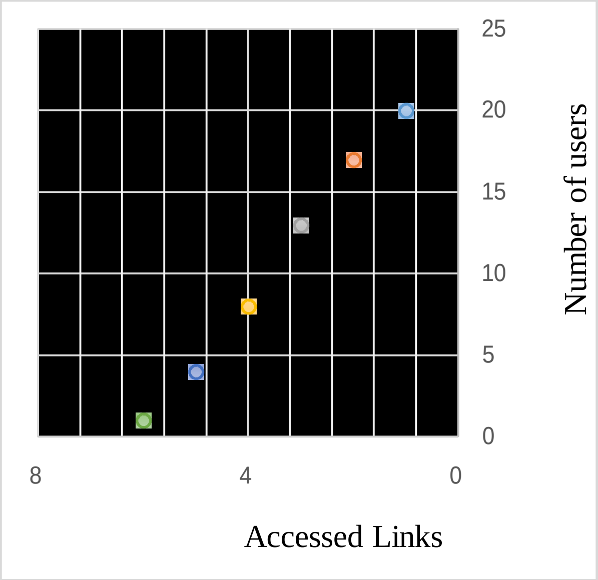

Figure 4.1: Accessed Links in Total

To evaluate the system performance, there are 20 users to test the accuracy of web page recommendations. The accuracy is calculated by the number of user's visited links individually divided by the total number of recommendation.

Scenario 2:- Applying apriori algorithm with 20% support, it gives the minimum links to the 20 users. It gives the minimum 2 links as recommendation to all users. The support shows that at least 20% of all accessed links should be visited by the users. It means at least 2 links should be visited by users. In this scenario all users have visited 2 links, 16 users visiting 3 links.

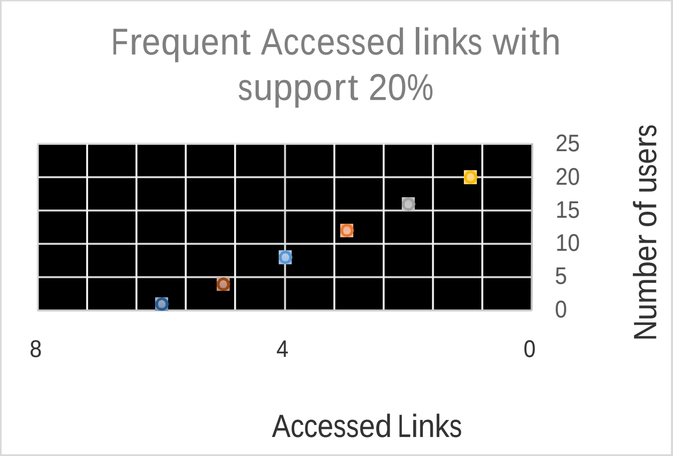

Figure 4.2: Frequent Accessed Links with Support 20%

Scenario 3:-Applying apriori algorithm with 40% support, it gives the minimum links to the 20 users. It gives the minimum 3 links as recommendation to all users. The support shows that atleast 40% of all accessed links should be visited by the users. It mean atleast 3 links should be visited by users.

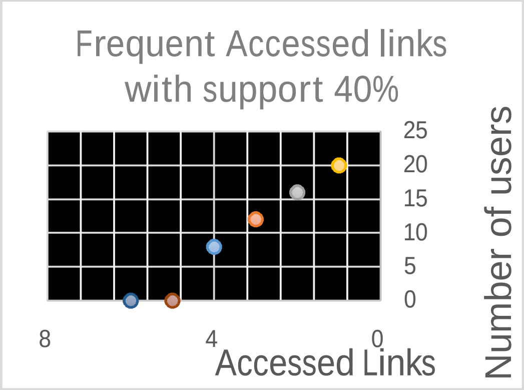

Figure 4.3: Frequent Accessed Links with Support 40%

Scenario 4:- Applying apriori algorithm with 40% support and with content mining which clustering the suers and visited links and with page rank, it gives the 3 links to the 20 users. It gives the minimum 3 links as recommendation to all users.

The support shows that atleast 40% of all accessed links should be visited by the users. It meanatleast 3 links should be visited by users. There are 6 clusters are formed as In this research have 6 links. And in this it gives more links to maximum users as compared to only with apriori algorithm recommendation system.

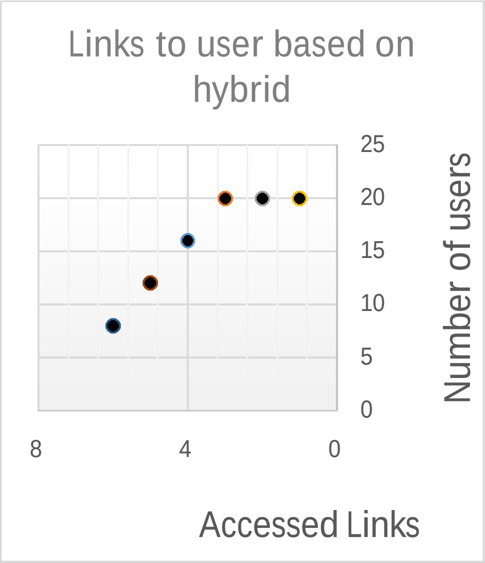

Figure 4.4: Recommended Links based on Hybrid with Support 40%

Scenario 5:- Usage mining based system gives no recommendation to new user but in hybrid with content mining and usage mining it gives 33.33% recommendation to new user also. This result is based on the clustering and page rank.

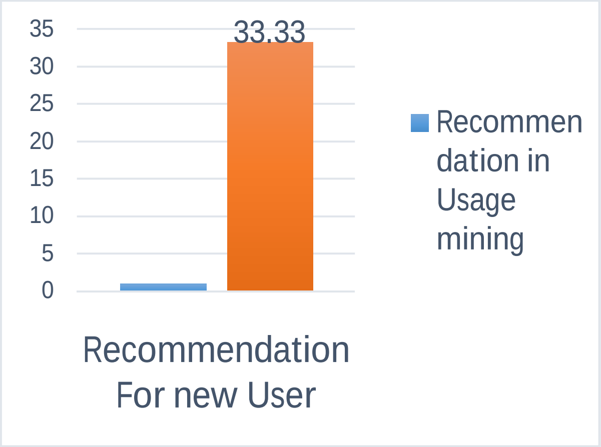

Figure 4.5: Recommended Links based on Hybrid for New User

Scenario 6:- The system based on hybrid it gives number of links to maximum users. Usage mining based system gives 30.05% average recommendation. While the Hybrid based recommendation system gives 50.05% average recommendation to users.

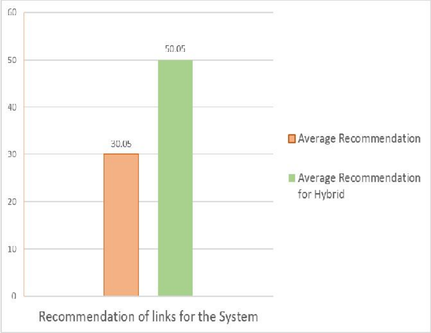

Figure 4.6: Average Recommended Links based on Hybrid and Usage Mining

V. CONCLUSION

In this research have incorporated the usage mining and the content mining for recommendation. In this research have used usage mining for the getting the user details like the web page visited, number of times visited, and date of visited, IP address. So it helps for sequencing the database of user history. Content mining is used for the content clustering. In this research have used DOM tree concept for getting the content of the page that user visited. The system gives the recommendation with the ranking of the page based on the number of times the page visited.

In this System it gives more recommendation with 40% support to maximum users. Also it gives 33.33%

recommendation to the new user. The average recommendation of recommender system based on the usage mining is 30% and the system based hybrid gives 50.05% recommendation.

REFERENCES

[1] Minxiao Lei, Lisa Fan Department of Computer Science, University of Regina,Regina, Saskatchewan , “A Web Personalization System Based on Users’ Interested Domains”, Proc. 7th IEEE Int. Conf. on Cognitive Informatics (ICCI'08) ©2008 IEEE

[2] Ford LumbanGaol Faculty of Computer Science Bina Nusantara UniversityIndonesia , “ Exploring the patterns of Habits of users using web log sequential pattern”, © 2010 IEEE

[3] Dario Vuljani, Lidia Rovan, MirtaBaranovi Faculty of Electrical Engineering and Computing,Croatia,“Semantically Enhanced web personalization Approaches and techniques”, ITI 2010 32nd Int. Conf. on Information Technology Interfaces, June 21-24, 2010.

[4] MatthewFredrikson University of Wisconsin, “Re-Imagining content personalization and in-browser Privacy”, pp.1081-6011/11

[5] KshitijaPol Datta Meghe College of Engineering,NitaPatil Datta Meghe College of Engineering,, “A Survey on Web Content Mining and extraction of Structured and Semi structured data”,pp.978-0-7695-3267-7/08 © 2008 IEEE

[6] Samira Khonsha Department of Computer Islamic Azad University, Zarghan Branch, YoungResearchers Club, Zarghan, Iran ,”New hybrid framework for web personalization”, 978-1-61284-486-2/111 ©2011 IEEE

[7] FBhushan Shankar Suryavanshi, NematollaahShiri, Sudhir P. Mudur Dept. of Computer Science and Software Engineering Concordia University, Montreal, Canada, “Fuzzy clustering approach for web mining”, 1 -4244-0363-4/06 ©2006 IEEE

[8] WANG Xiao-Gang Wuhan University of Science and Engineering, Wuhan City, Hubei Province, China 430073,”WebMining Based on User Access Patterns for Web Personalization”, 978-1-4244-4246-1/09©2009 IEEE