A Novel Web Page Recommender System for Anonymous Users Based on Clustering of Web Pages

Rajnikant Wagh
Research Scholar (Ph. D.)Dept. of Computer Engineering
R. C. Patel Institute of Technology,Shirpur Shirpur,
Maharashtra, India
rajnikantw@gmail.com

Prof. Dr. Jayantrao Patil
Principal & Professor
R. C. Patel Institute of Technology, Shirpur
Shirpur, Maharashtra, India
jbpatil@hotmail.com

***Abstract*— Information overload is major problem of today's internet use. Users frequently get much more information than needed. Web Personalization and recommender systems are becoming popular now days to overcome this problem. We have proposed a novel web page recommender system to improve browsing experience of anonymous users. We have used web usage mining technique for personalizing a web site and recommendation of web pages. This technique uses preprocessing, analysis for finding the relationship among web pages, clustering and classification phases of data mining. The preprocessing step aims at maintaining consistency in dataset. We have modelled the relationship among web pages with novel measures of distance matrix, occurrence frequency matrix and relationship matrix. A virtual graph is created corresponding to the relationship matrix to show the relationship among web pages. The proposed method partitions the virtual graph into various clusters i.e. navigation patterns by proposing an enhanced depth first search algorithm. It is a graph based partitioning algorithm. We classify the active user under consideration into one of the cluster by using LCS algorithm. Finally, we used a threshold value to recommend only optimum number of web pages. Use of novel measures for finding the relationship, use of threshold values at the time of formation of clusters as well as at the time of recommendation of web pages gives us better results in term of improved visit coherence, accuracy, coverage and F1 measures. We get max. 61% accuracy, 49.2% avg. coverage and 28.87% avg. F1 values in the recommendation of web pages. Similarly, we get 57.8% avg. visit coherence in the formation of clusters and a minimum of 15 % outliers.**

Keywords*— *Web Personalization, Recommender Systems, Web Usage Mining, Clustering, Classification

I. INTRODUCTION

The Continuous growth in the use of the internet imposes new methods for Information overload problem. Most web structures are complicated and large. It may mislead the users with unnecessary and unambiguous information. It is required to predict the user needs to improve surfing experience and providing them with what they want in less time. Web Personalization (WP) techniques or Recommender Systems (RS) are available solutions nowadays for this [1, 2].

Web Personalization is the process of customizing a Website to the needs of specific users taking benefit of knowledge acquired from the analysis of Web information i.e. content, structure, user profile data along with users' navigational behavior i.e. usage Data. Web personalization is a broader area covering recommender systems, adaptive Web sites, and customization. Customization process is done manually or semi-automatically whereas in Web personalization modifications in structure or content of Website are performed dynamically [3, 4].

In literature, Web personalization is also defined as the process of providing useful links, items, and objects to the user to save valuable time. Various data mining techniques are used for user analysis and recommendation purpose. It helps in improving the business or user satisfaction of various E-Commerce websites [5, 12].

II. LITERATURE SURVEY

There are three types of web mining approaches used to personalize a web. These are, web content mining, web structure mining and web usage mining.

Malik and Fyfe have focused a review of web personalization. The building blocks of web personalization viz. learning, matching and recommendations are discussed in detail with the recent trends. The challenges like high scalability of data, lack of performance, black box filtration, correct recommendation, and privacy issues are new opportunities for researchers [6].

Yang et al. developed a technique for personalizing Web page recommendation via collaborative filtering and topic aware markov model .They tried to predict the next request of pages that Web users are potentially interested in when surfing the Web. They implemented a graph-based iteration algorithm to discover users' interested topics, based on which user similarities are measured [7].

AlMurtadha et al. proposed an Improved Web Page Recommendation system using profile aggregation based on clustering of transactions. The authors have built recommendation system for anonyms' users or visitors' .For this purpose they assigned the current user to the best navigation profile with similar navigation activities [8].

Jalali et al. developed a recommendation system called WebPUM, an online prediction using longest common sequences algorithm (LCS) for classifying user navigation patterns to predict users' future intentions. To effectively provide online prediction, they proposed an approach for

classifying user navigation patterns to predict users' future intentions [9].

III. MOTIVATION AND PROBLEM STATEMENT

In view of the literature, review and practical limitations related to the use of content and structure mining for web page recommender systems, we have decided to develop a novel web page recommender system for anonymous users. Many techniques make use of content data, structure data, user profile data as well as usage data for achieving the goal of predicting user future requirements. Some systems use any combination of the aforementioned data for better results but with some limitations to accuracy and coverage of recommendations [10, 11]. Though there are many systems using the combination of content, structure, user profile and usage data, we have tried to focus our attention on usage data only. The reasons behind this are the practical limitations related to development as well as evaluating the system. Two to three options were available to us at the beginning of our research work. The first one was to develop our own website, host it over the internet and then use its user data from web servers for the better understanding of our website users and then personalize it in accordance with those users. In this case, we might have used all four types of data (content, structure, and user profile as well as usage data) for better understanding and better results. But this option had a limitation on the number of users or website visitors. Also, the second problem was a possibility of manipulating surfing patterns of visitors as per our requirements. The second alternative was to use already existing websites (especially E-commerce websites), use their log data and development of better algorithms for the recommendation of objects/items through our algorithms. But the problem with this technique was that no website owner was going to allow us for recommendations on their website. As well as it was really impractical to evaluate our system. These practical limitations led us to focus our attention on the sole use of web usage mining technique and remove other usable data (content, structure and profile data) from consideration. Finally, we decided to use weblog data of specific website. In our experiments, it is necessary to use such a dataset that allows us to analyse Web log data. Our experiments have been conducted on DePaul University CTI log file dataset (www.cs.depaul.edu) from web servers, pre-process it as per our needs, do the proper and better analysis, and develop our own algorithms or to contribute to existing algorithms for better results in terms of accuracy of recommendations.The proposed system aims at improving the browsing experience of anonymous users with the following objectives and steps.

IV. RESEARCH CONTRIBUTIONS

Following are the novel contributions of the proposed research work:

- Development of new relationship measures based on distance among requests of web pages in sessions as well as occurrence frequencies to get appropriate relationship among web pages. These are not developed or proposed until now in any web page recommendation system.
- The proposed measures/ implementation try to give justice to all the web pages of a website. It considers all the sessions (out of total 13745 sessions) where both web pages have occurred together. The measures are developed in such a way that they create normalized values between zero and one.
- Preparation of virtual graph corresponding to relationship matrix. The nodes of the graph are the web pages of a website whereas the edges between the nodes represent the relationship value obtained by distance and occurrence frequency measures.
- Partitioning of a graph into clusters with the enhanced depth-first search algorithm. The clusters prepared are nothing but the representatives of navigation patterns followed by many users in the past. While partitioning, we used threshold values of edges as well as threshold values of cluster size as main additional parameters to get more effective navigation patterns.
- The longest common subsequence algorithm does classification of active users. It maps the active user to one of the cluster found earlier.
- Recommendation of remaining web pages of that cluster. For this, we first ranked the web pages according to the final weight value, applied threshold values, and recommended only those web pages, which fulfills threshold criteria.
- The proposed technique is tested on CTI dataset w.r.t. various performance measures like Visit Coherence, Number of Outliers, Number of Clusters, Accuracy, Coverage, and F1.

New distance measure, occurrence frequency measure as well as relationship measure helped to find the most appropriate relationship among the web pages. This is further improved by the application of enhanced depth first search algorithm as well as longest common subsequence algorithm. The basic features of web page recommendation like maximized accuracy and maximized coverage are accomplished successfully.

V. SYSTEM DESIGN

Based on previous research findings in data mining and web mining, we decided to follow a specific path towards the design of our system. We have developed a new architecture for predicting user future requests. The model is partitioned into two different phases, offline and online. Though separated in the model, the offline phase strongly affects the online phase. The offline phase is used to extract user navigation sessions from the original Web user log files. An enhanced clustering algorithm based on graph partitioning is introduced for navigation patterns mining. An online phase within the prediction engine has the task to predict user future requests. Classifying the user current activities based on navigation patterns in a particular Web site is the main objective of this phase. In addition, creating a list of recommended Web pages as a prediction of user future movement is another objective in this phase. The main online component is the prediction engine.

5.1 Offline Phase of Proposed System:

This phase consists of two main modules, which are data pre-processing and navigation pattern mining.

Identify applicable funding agency here. If none, delete this text box.

5.1.1 Data pre-processing module: It is designed to extract user navigation patterns from the original Web user log files. An enhanced clustering algorithm based on graph partitioning is introduced for navigation patterns mining. The pre-processing step discuss about what type of log files we collect and how we get them converted to the proper final filtered pre-processed dataset. **1) Data collection:** At the beginning of data pre-processing, Web server log files are collected from several Web servers to be utilized in various tasks. **2) Data cleaning:** This step is performed to eliminate the irrelevant entries from the log files, which includes Web robots, spiders and crawlers, picture files associated with requests for particular pages, CGI files, and other irrelevant files. **3) Data structuration:** After cleaning the Web user log files, a series of tasks are applied on the cleaned dataset to identify users and sessions.**4) Data Summarization:** After structuring the data, data file is transferred to a relational database. Data transformation and aggregated data computation for user sessions is applied after this.

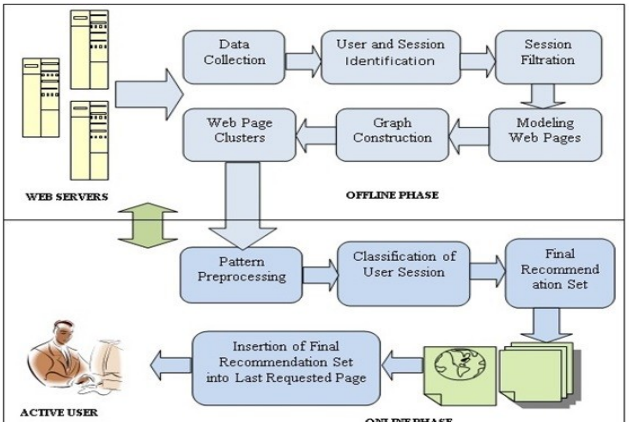

Figure 1: Reference Framework for proposed work

5.1.2 Navigation Pattern Mining: In this work, a clustering model is used for navigation pattern mining. The model exploits the graph-partitioning algorithm by applying a new method for creating an undirected graph. The clustering model is built to find a collection of related pages at a particular Web site, relying on the visit-coherence assumption. The pages that a user visits during one interaction with the site tend to be conceptually related. The process of the clustering takes three steps. **Step 1:** Generation of Relationship Matrix for Web Pages of a Website. **Step 2:** Creation of Weighted Graph with respect to Relationship Matrix. **Step 3:** Partitioning of Graph and Formation of Clusters.

We know that the web pages are requested in a specific order while surfing. Taking the benefit of same surfing strategy, we have decided to develop some mechanism to establish a connection among surfing patterns of various users. We have used all user sessions of a dataset. Every user session contains a sequence of pages in accordance with three fields' viz. time stamp, URL visited and referrer. We use this input data to find the relationship among web pages. Relationship Matrix is a combination of two measures i.e. Distance Matrix and Occurrence Frequency Matrix. We use below process to find Distance Matrix, Occurrence Frequency Matrix, and Relationship Matrix.

Distance Matrix: We proposed a new measure of finding the distance between requests of every two pages in a session. Distance Matrix represents the distance of requests for every two pages in a session.

$$DM_{x,y} = \text{Avg.} \sum_{i=1}^{n} \frac{|d(X_i) - d(Y_i)|}{\text{Session size i of unique pages}}$$

Where d (xi) is the position of webpage x in i^{th} session, d (yi) is the position of webpage y in i^{th} session. We consider all the sessions where webpage x and webpage y both have occurred together. It is the indication of request difference between two pages x and y in session i. If the difference between the request time of page x and page y is more, we mean that the pages x and y are less alike to be requested in a session and vice a versa. Concluding it in another way, the more the value of DMx, y, the less they are related whereas the less the value of DMx,y means the more they are related. The formula is normalized so that all the values for the Distance Matrix are between 0 and 1.

Occurrence Frequency Matrix: We also learned from the previous finding that the session can be considered as useful if both pages x and y are occurring in that session. Taking the same thing as a key, we have decided to use all the sessions where both pages x and y occur simultaneously. Occurrence Frequency measures the average occurrence of both pages in each session as shown below.

$$FM_{x,y} = \frac{T_{xy}}{\text{Avg.}(T_x, T_y)}$$

Where Txy= Total number of sessions where page x and page y both occur together, Tx = Number of sessions where only page x occurs, Ty=Number of sessions where only page y occurs. By considering the average occurrence of both pages, we try to give equal importance to the occurrence of both pages in a session. This formula is also normalized so that all the values for Occurrence Frequency Matrix are between 0 and 1.

Relationship Matrix: Distance Measure and Occurrence Frequency Measure are two strong indicators of connectivity among each pair of web pages. Therefore, in Relationship Matrix, Distance and Occurrence Frequency are valued equally. We use the harmonic mean of Distance Matrix and Occurrence Frequency Matrix to represent the relationship between two pages. We use this Relationship Matrix for the creation of a weighted undirected graph in the next step.

$$RM_{x,y} = \frac{(2 * DM_{x,y} * FM_{x,y})}{DM_{x,y} + FM_{x,y}}$$

Relationship Matrix tries to approximate the visits among web pages. It is the indication that more the value of RM for any two pages x and y, they are more closely related. These two pages are requested frequently one after the other in more sessions. We use this concept for the creation of weighted undirected graph in next step, where nodes of the graph are web pages of a website and the edges between these pages represents the relationship found between these web pages in the previous step.

5.2 Online Phase

This phase generates list of recommended web pages to the active user. It utilizes the clusters of navigation patterns generated out of offline phase. It predicts user future requests beforehand and produce short-term view of potentially useful links. The proposed method inserts the links of recommended Web pages in the last requested Web page by

the active user. For this, we follow below mentioned steps. **Step 1:** Preprocessing of user active session as well as navigation patterns. **Step 2:** Classification of active user by LCS algorithm. **Step 3:** Create and recommend a set of Web pages.

VI. EXPERIMENTAL EVALUATION

We have implemented our work on DePaul University CTI log file dataset (*www.cs.depaul.edu*). This data set contains the data for the main DePaul CTI Web server (http://www.cs.depaul.edu).

TABLE 1: EXPERIMENTAL EVALUATION

Edge Threshold	Cluster	Outlier	Visit Coherence	Accuracy	Coverage	F1
0	1	0	100	0	100	0
0.1	1	0	100	0	91	0
0.2	28	14	89	8	84	14
0.3	69	24	58	31	55	41
0.4	80	41	54	49	48	49
0.5	71	59	47	62	35	46
0.6	60	64	41	44	23	28
0.7	31	73	37	30	21	25
0.8	24	86	33	21	18	18
0.9	8	91	31	17	16	16
1	0	100	0	0	0	0

The data is based on a random sample of users visiting this site for a 2-week period during April of 2002. The original (unfiltered) data contained 20950 sessions from 5446 users. The filtered data files are produced by filtering low support page views and eliminating sessions of size 1. The filtered data contains 13745 sessions and 683 page views. Two experiments are conducted using above-mentioned implementation specifications. In the first experiment, after modeling of Web pages clustering of Web pages is done using enhanced depth first search algorithm. In the second experiment, user active session is classified in one of the cluster using LCS algorithm. Table 1 summarizes experimental results of proposed research work.

VII. CONCLUSION

We have developed a novel and improved web page recommender system for better surfing experience of users. The proposed technique is based on finding appropriate relationship weights among the web pages of a web site. The modeling among the Web pages is done by measuring the distance as well as occurrence frequency. The enhanced clustering done on this relationship matrix helped to form appropriate clusters. We classified the active users using LCS. The Threshold used in the last phase of recommendations improves the accuracy of web page recommendation system. Improved accuracy of 61% in recommendation of web pages to active user definitely saves the surfing time of user. The users will try not to visit the web pages of a site, which are less relevant. We have also got around 37 % of coverage for largest accuracy value. We get max. 61% accuracy, 49.2% avg. coverage and 28.87% avg. F1 values in the recommendation of web pages. Similarly, we get 57.8% avg. visit coherence in the formation of clusters and a minimum of 15 % outliers. Semantic knowledge about the underlying domain may improve quality of recommendations further.

ACKNOWLEDGMENT

We are thankful to North Maharashtra University Jalgaon for providing assistance and support for successful completion of project under Vice Chancellor Research Motivation Scheme (VCRMS).

REFERENCES

[1] M. Eirinaki and M. Vazirgiannis, (2003), "Web Mining for Web Personalization", In Proceedings of ACM Transactions on Internet Technology (TOIT).ACM, Athens, Greece, 3(1), pp.1-38, http://doi.acm.org/10.1145/643477.643478

[2] G. Adomavicius and A. Tuzhilin, (2005), "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions", IEEE Transactions on Knowledge and Data Engineering, 17(6), pp.734–749.

[3] U. Gulden and M. Matthew, (2008), "Personalization Techniques and Recommender Systems: Series in Machine Perception and Artificial Intelligence", World Scientific Press, Vol. 70, Singapore,

[4] M. Bamshed and A. Sarabjot Singh, (2005), "Intelligent Techniques for Web Personalization", Springer, New York.

[5] R. Francesco, R. Lior, and S. Bracha, (2015), "Recommender Systems Handbook", Springer, New York.

[6] Z. Malik and C. Fyfe, (2012), "Review of Web Personalization", Journal of Engineering Technologies in Web Intelligence, 4(3).

[7] Q. Yang, J. Fan, J. Wang, and L. Zhou, (2010), "Personalizing Web Page Recommendation via Collaborative Filtering and Topic-Aware Markov Model", IEEE International Conference on Data Mining, 1(1), pp. 1145-1150.

[8] Y. AlMurtadha, N. Sulaiman, N. Mustapha and N. Udzir, (2011), "IPACT: Improved Web Page Recommendation System Using Profile Aggregation Based On Clustering of Transactions", American Journal of Applied Sciences, 8 (3), pp. 277-283.

[9] M. Jalali, N. Mustapha, N. Sulaiman and A. Mamat, (2010), "WebPUM: A Web-based Recommendation System to Predict User Future Movements", Expert Systems Applications, 37, pp. 6201-6212.

[10] H. Liu and V. Keselj,(2007), "Combined Mining of Web Server Logs and Web

[11] Contents for Classifying User Navigation Patterns and Predicting Users' Future Request", Data and Knowledge Engineering, 61(2), pp.304-330.Conference Short Name:WOODSTOCK'18

[12] B. Mobasher, R. Colley, and J. Shrivastav, (2000), "Automatic Personalization based on Web Usage Mining", Communications of the ACM, 43 (8), pp. 142-151.

[13] C. Sumathi, R. Valli and T. Santahnam,(2010), "An Application of Session Based Clustering to Analyze Web Pages of User Interest from Web Log Files", Journal of Computer Science, 6(1), pp.785-793.