https://dx.doi.org/10.24001/ijaems.3.2.18 *ISSN: 2454-1311*

An Approach to Recommend Pages to User after Path Completion

Ms. Priyanka Makkar, Ms. D.D Kadam

Department of Computer, Pune University, Pune

Abstract— *With the huge amount of data available on the web, the effective and timely search is most important aspects. Because of tremendous usage of web the size of the log files are becoming huge day by day. Web Mining which is an application of data mining can be used to improve web performance by generating patterns in the log files. Web Pre-fetching based on recommendation can be used to improve web performance. This paper proposes an architecture that recommends pages to the user based on studying the browsing activity of the user and other user of same usage profile and completing it in every aspects and then recommending the pages which can be pre-fetched to improve web performance. This paper proposes about the path completion which shows users interest and then recommending pages to the user which will result in effective recommendation.*

Keywords— Web Usage Mining, Web recommendation, K-Means clustering, path completion, Perfecting, web log file, pattern discovery.

I. INTRODUCTION

Since a huge amount of data is present on the web and as it is growing at a tremendous rate day by day so retrieval of effective information and that too on time is very important and hence the topic of research among researchers. Web usage mining can be used to provide effective and timely information to the user. Web usage mining is used to mined information from web access log, Web logs contains information about access pattern of the user which can be studied and according to the analysis we can recommend pages to the user to help increase web efficiency and therefore web usage mining is a process of extracting pattern of interest from their access pattern of web pages. Web usage mining consists of three main steps: Data preprocessing, Pattern Extraction, and Pattern Analysis.

Pre-processing is an important process and since the Web structure is very complex and almost 80% mining process is done at this phase. It includes cleaning, session identification, and path completion.

The second step in web usage mining is pattern extraction in which data mining algorithms like association rule mining techniques, clustering, classification etc are applied to pre-processed data[1]. The third step is pattern analysis in which we present information into knowledge.

In this paper we are studying logs and completing paths in the logs using petrinets . Petri Nets (PN) is a high-level graphical model widely used in modeling system activities with concurrency.PN can store the analyzed results in a matrix for future follow-up analyses. According to the definition it is formally defined as a 5-tuple PN of (P, T, I, O, M0), where

(1) P = {p1, p2… pm}, a finite set of places;

(2) T = {t1, t2… tn}, a finite set of transitions; P□ T≠Ø, and P∩T= Ø;

(3) I: P x T→N, an input function that defines directed arcs from places to transitions, where N is a set of nonnegative integers;

(4) O: T x P →N, an output function that defines directed arcs from transitions to places;

(5) M0: P →N, the initial marking. A marking is an assignment of tokens to a place[10];

PN is carried out by firing transitions. A transition, t, is said to be enabled if each input place, p, of t contains at least an amount of token equal to the weight of the directed arc connected to t from p. In a PN model, we can utilize the different token amounts in the places to represent the different system states. Since a fire of transition in the system often can be associated with a change of the token amount in a place, PN hence can represent, or model, the system dynamic behaviors via the fire of transitions. An incidence matrix records all token-amount changes in all places after all fired transitions. For PN with n transitions and m places, the incidence matrix A, where A=[aij], is an n×m matrix of integers; its typical entry is given by

aij = aij+ - aij- where aij+= O (ti, pj), the weight of the arc from Transition i to its Output Place j, and aij-=I (ti, pj), the weight of the arc to Transition i from its Input Place j; aij+, aij- and aij represent the number of tokens removed, added, and changed in Place pj, respectively, when Transition ti fires once.

During the processing or operations, this method can also simultaneously trace out what are the possible intermediate states during the transitions from the initial state to the destination one. In a PN model, a marking Mi

https://dx.doi.org/10.24001/ijaems.3.2.18 *ISSN: 2454-1311*

is said to be reachable from a marking, M0, if there exist a sequence of transition firings which can transform a marking, M0, to Mi [10].

In this paper, a model is proposed in which we recommend user pages based on their access pattern and also the list of pages visited by different users having similar interest or users with similar profile. The rest of this paper is organized as follows: In section 2 , we discussed about related work. Section 3 presents the proposed model for the recommendation and also its explanation. In section 4 we present the algorithms explaining about proposed system. Finally, section 5 concludes the paper with the direction for future work.

II. RELATED WORK

There are various recommendation systems proposed by different researchers. This type of recommendation system is used to predict the user navigation behavior and their preferences using web log data.

Bamshad Mobasher [2] presented a system called Web personalizer which provides dynamic recommendations, as a list of hypertext links, to users. In preprocessing phase, the data mining techniques (i.e. clustering, sequence pattern discovery and association rules) are used to obtain the aggregate usage profiles.[3].

the Recommender systems based on the user's access patterns using model based clustering and recommendations has been provided as per user requirement.[4].AlMurtadha et al. [5] have focused on improving the prediction of the next visited web pages and recommends it to the current anonymous user by assigning them to the best navigation profiles obtained by previous navigations of similar interested users. Mrs. V. Chitraa[6] discussed about various steps involve in data preprocessing.The process of session clustering was explained by Li Chaofeng [7]The advantage of path completion was explained by Nirali Honest [8].and Dr. Atul Patel Dr. Bankim Patel "A Study Of Path Completion Techniques In Web Usage Mining"[8].

III. PROPOSED MODEL FOR RECOMMENDATION

The proposed work consist of two main parts, namely front end and back end as shown in fig:1.

In the front end the three main steps are taken into consideration. First is to update the knowledge base as per request , secondly check its similarity with knowledge base and last is to recommend it based on similarity measures.

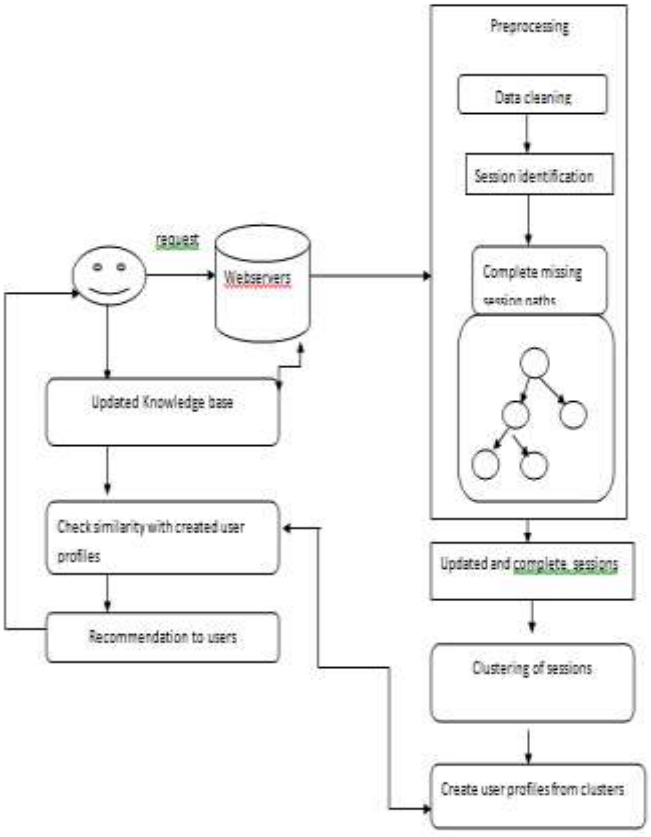

Fig. 1: Proposed Recommendation system

In the back end we are dealing with four steps ,First is preprocessing of logs ,which includes completion of paths and cleaning of logs, session identification based on time. Next is to create a updated logs .It is then partitioned into clusters of user navigation with similar browsing activity using K-means clustering algorithm.Next is to create profiles based on clusters, The usage profile contains only those web pages that passed certain confidence support and weights values The usage profile is constructed as a set of page view and the same knowledge is used for recommendation to current user.

3.1 Data preprocessing

It is the foremost phase before the logs can be used to obtain useful information. It consist of following step:

1. Data Cleaning
2. User and Session Identification
3. Path Completion

3.1.1 Data Cleaning

Data cleaning is a process of removing irrelevant data from the logs. In data cleaning the records with .GIF, .JPEG extensions are cleaned from the logs .Moreover records with failed status code i.e status code with less

https://dx.doi.org/10.24001/ijaems.3.2.18 *ISSN: 2454-1311*

than 200 and greater than 299 are removed from the logs. Pages with timestamp less than 2 sec are also removed from the logs.

3.1.2 User and Session Identification

User Identification:Records with different IP address are considered as different users else if IP address are same then the browser and operating system information is checked in the user agent field ,if it is different then they are considered as different users.

Session Identification:A session is a sequence of page accessed by a user during visit to the site. The aim of session identification is to divide the page accesses of each user into different sessions. One way to identify sessions is based on time , The set of pages visited by a specific user in 30 minutes by R.Cooley [9] is assumed to be in one session and if it exceeds the time stamp of 30mins it is considered to be in different session. Sessions are also identified based on page stay time which is evaluated based on difference between two timestamps. If it exceeds 10 minutes then it is considered as in new session.

3.1.3 Path Completion

Path completion step followed by session identification. There are many incomplete paths in the user session , those incomplete paths are removed if they are not handled properly. So path completion is important step in the entire process. Incomplete paths are due to pages accessed through proxy servers or cached versions of the pages are used by the user , So path completion step is carried out to identify missing pages and missing path are completed. Petri nets is one of the algorithms that can be used for path completion.

3.2 Session Clustering and user profiles creation

The next step is to identify similar access pattern from the user session file. Each cluster consist of many sessions of "similar" access patterns using K-Means algorithm[7] and Weighted K-Means algorithm [12]. Finally, user profiles are generated based on the clusters. The usage profile contains only those web pages that passed certain confidence support and weights values. The confidence support determines the frequent occurrence on those pages in the cluster. These profiles don't consider specific users.

3.3 Similarity checking and recommendation

In the front end when the user accessed any page its similarity is checked with generated profiles. Similarity is determined using the well-known Cosine similarity measure and Hamming Similarity and recommended list of pages is created whose similarity values is greater than the threshold and finally recommended to the active users.

IV. ALGORITHM FOR RECOMMENDATION SYSTEM

Algorithm1: knowledgebasecreation(query i+1)
{
1.Make entry in logs querylog =querylog U queryi+1 .

2. Preprocesslogs(querylog)
3.discover patterns(queryi+1)
4.Create user profile and update in knowledge base.
}

Algorithm2: Preprocesslogs(querylog)
{1. Querylog=clean logs(querylog)
2.Identify sessions (Querylog)
3. if (web Site Structure exists)
break;
else
create structure()
end if
4. return
}

Algorithm3: Recomendation (qi+1)
{
1. for every new query
2 .s= similariy(qi+1)
3. if(s>threshold)
return URLS
else
break;
}

V. CONCLUSION

This paper has focused on improving web logs by path completion and then recommends pages based on user access pattern and also access pattern of the different user of similar usage profile. In this paper we have generated user profile by clustering pages and thereby recommending pages to users which can be cached and can reduce the access time of user and thereby improve web performance

REFERENCES

[1] Mrs. V. Chitraa and Dr. Antony Selvadoss Thanamani "A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing", International Journal of Computer Applications (0975 – 8887).

https://dx.doi.org/10.24001/ijaems.3.2.18 *ISSN: 2454-1311*

[2] Bamshad Mobasher, 2001. "Web Personalizer: " A Server Side Recommender System Based on Web Usage Mining" .Technical Repor TR01-010, SchooL of Compute Science, telecommunications and Information Systems, DePaul University, Chicago, IL, USA.

[3] R.Thiyagarajan, K.Thangavel, R.Rathipriya, "Recommendation of Web Pages using Weighte d K- Means Clustering " International Journal of Computer Applications (0975 – 8887) Volume 86 – No 14, January 2014.

[4] R. Padmaja Valli, Sumathi ,C.P. Sumathi ,"Automatic Recommendation of Web Pages in Web Usage Mining ",International Journal on Computer Science and Engineering Vol. 02, No. 09, 2010, 3046-3052.

[5] Al Murtadha, Y.M., M.N.B. Sulaiman, N. Mustapha and N.I. Udzir ,2010 ." Mining web navigation profiles for Recommendation system . Inform.Technol.J.,9: 790-796 DOI:10.3923/itj.2010.790.796

[6] Mrs. V. Chitraa and Dr. Antony Selvadoss Thanamani "ANovel Technique for Sessions Identification in Web Usage Mining Preprocessing", International Journal of Computer Applications (0975 – 8887).

[7] Li Chaofeng "Research based on Web Session Clustering" Journal of software, VOL. 4, NO. 5, July 2009.

[8] Nirali Honest and Dr. Atul Patel Dr. Bankim Patel "A Study of Path Completion Techniques In Web Usage Mining", 2015 IEEE International Conference on Computational Intelligence & Communication Technology.

[9] Cooley, R., B. Mobasher and J.Srivatsava, 1997 "Web mining information and pattern discovery on the world wide web" Proceeding of the 9th IEEE International Conference on tools with Artificial Intelligence.

[10] Po-Zung Chen, Chu-Hao Sun, Shih-Yang Yang, "Modeling And Analysis the Web Structure Using Stochastic Time PetriNets", Journal of Software, Vol. 3, No. 8, November 2008