AN EFFECTIVE RECOMMENDATION SYSTEM FOR TREASURING WEB USER’S BROWSING TIME

K.S.Sakthi Priya[1], M.S.Bhuvaneswari[2], K.Muneeswaran[3]
[1]PG Student, [2]Assistant Professor, [3]Senior Professor

Abstract

Nowadays, the count of web user’s keeps on increasing. Therefore, all the business organizations start to focus on constructing own website in order to attract the new customers and to retain the existing customers. Mining web users’ interest has become increasingly important to construct an optimized website. Weblog file plays a vital role in mining web users navigational and accessibility pattern. The recommendation system is the system which recommends the active user a page that may be interested by them. In order to construct an optimized website, two new techniques have been proposed in this proposed system. One is indexing technique which is used for fast retrieval of page to be recommended. The second one is construction of weighted web user session which is a sequence of pages visited by users along with additional information calculated by considering the frequency of visit by the user to the page that reflects the user’s interest on the visited page. This weighted session is used to identify exact similarity between two sessions and also used to find user interested page exactly and accurately compared to existing recommendation system. Sliding window technique is used to slide the window with *w* recently visited pages which is very useful for construction of inverted indexed page table. This paper involves two stages –First one is an offline stage which involves Data preprocessing and Inverted Indexed Page Table construction. The second stage is Online Recommendation Engine. An experiment is also carried out in this to find an optimized *w* value which provides highest accuracy. Finally, a page interested by the active user is quickly and accurately recommended by the proposed system.

Keywords: Web server logs; weighted Session; Indexing technique; Recommendation System; Inverted Indexed Page Table Construction; scalable KNN approach;

I. INTRODUCTION

Web, as per the name, it is the place where a huge amount of useful data about the web user’s gets accumulated every day. This data is further enriched by the data engineers through data analyzing techniques which will provide a win-win business situation. Web Mining is a type of data mining technique which will enable the business organization to obtain further business opportunities for companies by extracting and processing the useful data available on the World Wide Web. Web mining is divided into three types based on the type of data to be mined. The first type is content mining which is the process of mining needed information from web documents. The second type is structure mining which is the process of mining needed structural information from the World Wide Web. The third type is usage mining, which is the process of extracting user access pattern from server logs [3]. Web usage mining is a type of data mining technique which mine’s the information about the web usage pattern of each web user [4]. Numerous web usage mining techniques are proposed every year which is used to extract web user’s navigational pattern and web usage behavior in order to design an optimized website which satisfies the end users [5]. Weblog file which records each web user’s navigational activities plays vital role in web usage mining [2]. In [15], the authors proposed an indexing technique which is applied to retrieve the page interested by the active user quickly and accurately by constructing Indexed inverted page table. They also used sliding window technique in their work to slide the recently visited page and to construct inverted indexed page table. There are two types of recommendation based on the type of data the system consider, in order to calculate similarity for recommending interested page to the active user. One is collaborative filtering system and another one is content based filtering system [8]. The collaborative filtering system recommend page by finding correlation between two users by using some explicit measures. Unlike collaborative filtering system, content based system recommends page by considering the correlation between the content of the page and the user's preferences. Compared to content based filtering system, collaborative filtering have much scope on research and easy to construct, since it needs only web log file as an input to construct an recommendation model. In this work, we have used two major techniques to overcome the problem of existing recommendation system. Major part of content based filtering system is to find similarity. Therefore, inorder to find exact similarity between two sessions, the session is reconstructed into augmented sessions which reflect the user's interest on the visited sequence of page.

The rest of this paper is organized in following sections: In section 2, the existing relevant literature on recommendation systems done in the context of the proposed system are discussed. Section 3 provides the detailed design of the proposed system which involves the data preprocessing, user identification process, the inverted indexed page table construction and finally, a techniques adopted to build a recommendation system is discussed. Section 4 discusses the recommendation system validity measure which is used to validate the resultant recommendation model. Section 5 contains the experiments that were carried out and the results obtained. Section 6 concludes the paper with future enhancement.

II. RELATED WORK

The main task of data preprocessing is to clean the raw web log file in order reduce size of the data since all the data obtained from the web log file is not needed for constructing their proposed model [13]. In Reference [10], the authors proposed a novel data cleaning, user and session identification techniques. The various data sources for collecting usage information were also discussed. In Reference [12] the authors discussed an advanced data cleaning algorithms to construct a user session from the web log file for effective mining. In Reference [11] the author discussed various heuristics methods to generate the sessions from web server log file. Time-Oriented heuristics is widely used conventional method with varying time limits. Time limit of 30 minutes is proved to be effective for generating expected result and also used in many business projects. In Reference [7], the authors proposed an enhanced session identification technique which augments additional information to the normal session that reflects the user's interest on the visited page. This novel session identification technique constructs session by calculating the relevance of user on a visited page. In Reference [9], the author used a KNN approach in his work to recommend a user interested page by comparing top k sessions with active user session. Predicting online user navigational pattern by using clustering technique are not accurate and efficient as KNN approach. So he proposed this approach in his work in order to build an optimized recommendation system.

Indexing technique which is used to construct index table in textbook, is used in our proposed work to retrieve the needed page quickly. This indexing technique is used to reduce the response time of the proposed recommendation system. Another approach used in this proposed work is constructing weighted session. To find exact similarity between active user session and previously logged sessions, a session with sequence of pages visited by the user is not enough, a session with sequence of page along with values which depicts the user's real interest on the visited page is needed. As a solution, identified sessions are reconstructed into weighted sessions which reflect the users' exact interest on the visited page. Even though the KNN approach is efficient compared to the clustering technique for predicting online pattern, it suffers from scalability problem. To overcome this problem, a scalable KNN approach with inverted indexed page table is used in our system.

III. SYSTEM METHODOLOGY

This system consists of three modules, A)Data Preprocessing - Data Cleaning and User identification, session identification B)Inverted

Indexed Page Table construction C)Recommendation system which involves three stages – reconstruction of normal session into weighted session, computing similarity between subset of session and active user session and finally, a top k similar session is selected to recommend a set of f pages which occurs frequently after the sequence which is similar to the active user sequence.

A. Data Preprocessing

Weblog file is a set of records which automatically created and maintained by a web server. Every request made by the user to the Website, including each view of an HTML document, image or other content, is recorded as a new entry. Each record in the file is essentially one line of text for each hit to the website. This contains information about who was visiting the site, where they came from, and exactly what they were doing on the website. Weblog file could be collected at the server side, client side and proxy servers. College web server log is taken as an input for this work which is in extended log format (ELF). The Extended log format is the Standardized text like format used by the web server when generating the log files like the common log format but it gives more information and flexibility because it contain 21 fields.

i) Data Cleaning

In this process, the record with irrelevant data is removed which are not significant for constructing recommendation system. Record with a multimedia request (.mpg, .jpeg, .jpg, .gif etc), record with unsuccessful user request, record with a request method other than GET and record with missing data ('-') were removed. Record with robot and anonymous request are also removed because processing such records doesn't provide any significant information. Finally, after removing the irrelevant data, tokenize the records into fields and store it into the database for further process.

ii) User identification

User identification is the most crucial step for constructing recommendation model since the similar user for active user need to be identified in order to recommend page which may be interested by the active user. In this work, the unique users are identified from the combination of hostname, OS and browser field extracted from cleaned log file in common log format. Each unique combination represent unique user. User identification is essential for next session identification process.

iii) Session identification

Session identification is the second most crucial step in web usage mining since it gives a structured format derived from unstructured raw web log file. The sequence of pages requested by an individual user is known as session. It is the unit of interaction between a web user and the server. Time-Oriented Heuristics method is used in this work to identify the session and the resultant session would be in the form $S_i = \{P_1, P_2, P_5\}$. It uses a maximum time limit to split the records into sessions (i.e.) a threshold (30 min) is fixed for each session. While adding a new log record to a session, if the total duration of the session exceeds this threshold and number of records in the session exceeds 30, then the log record is put in a new session.

B. Inverted Indexed Page Table Construction

Existing recommendation systems uses KNN approach which is applicable only for weblog file with small size. Once the size of the file get increased, then this KNN approach is not applicable. In order to overcome the non-scalability problem in KNN approach, a scalable KNN approach by constructing inverted indexed page table is used. This inverted indexed page table is used to speed up the searching process, which retrieves only the subset of sessions from the previously logged sessions that is similar to the current active user browsing sequence. By using indexing technique, it need to compare the active user session with only subset of previously identified session instead of comparing the active user session with all previously logged sessions. Sliding window technique is also adopted in our work to construct inverted indexed page table. While an active user browsing the website, the pages recently viewed by them are tracked. Each time the visitor requests a new page, slide the window by one. Consequently, the newly requested page will be added and the oldest page in the window will be dropped in the active user sequence. Assume *w* as a size of sliding window, so that recommendation system need to wait until the active user reaches the *w* sequence of visited page in that website. Constructed Inverted Indexed page table should contain *w* sized possible keys generated from previously logged user sessions with list of session id's which contain that key. Finally, an inverted indexed page table is outputted from this module which is given as an input to the online stage of the proposed recommendation model. The sample sessions with sequence of pages visited is given in Table I and the sample inverted indexed page constructed from the sample sessions is given in Table II. Here, consider *w = 3*.

Table I
Sample Sessions with sequence of pages

Session ID	List of Visited Pages
S1	1,2,3,4,7
S2	3,4,7,8
S3	1,5,7,9,11
S4	2,3,4,7,8
S5	1,2,3,5,7,9,11

Table II
Sample Inverted Indexed Page Table

Possible Keys	List of Session ID's
1,2,3	S1,S5
2,3,4	S1,S4
3,4,7	S1,S2,S4
4,7,8	S2,S4
1,5,7	S3
5,7,9	S3,S5
7,9,11	S3,S5
2,3,5	S5
3,5,7	S5

C. Recommendation System

In order to recommend exactly the user interested page, an optimized recommendation system is constructed that involves three stages which is explained as follows.

i) *Weighted Session construction*

First and Foremost stage in our proposed recommendation system is reconstruction of normal session into weighted session. Sessions identified in the offline stage is not much sufficient for grouping users' of similar interest. Sometimes two different users' visit the same sequence of page on the website but their interest on the visited page may vary. So grouping web user by taking into consideration the sequence of pages visited by users may lead to false categorizing of web users' of common browsing behavior. The web user's habits, interests and expectations should also be considered for grouping users of similar interest by measuring the relevance of pages in every session. The weighted session is the session with additional information which reflects the users' interest on the visited page. Construction of weighted session involves finding user's interest on visited page. The user's interest on the page is calculated by finding TF-IDF value for each page visited by the user. Most well-known technique in information retrieval is TF-IDF value calculation combined with cosine similarity to find similar session to the current active user session. TF-IDF formula which is popularly used in information retrieval journal is given in equation (1).

$$\mathrm{TF}-\mathrm{IDF} = \frac{\log\left(1+\dfrac{n(s,p)}{n(s)}\right)}{n(p)} \qquad (1)$$

Where n (s, p) is the number of occurrences of page p in session s, n (s) is the number of pages in session s. n (p) is the number of sessions containing page p. TF-IDF value is calculated by considering the frequency property of the page. Finally, the normal sessions are added with additional information and converted into weighted session. The main aim of this reconstruction is to find exact similarity between the previously logged users' sessions and the current active user session. The weighted session is outputted from this stage and given as an input to the next stage.

ii) User's Interest based similarity measure for computing similarity between subset of previously logged users' session and current active user session

Similarity measure is used to capture the resemblance between two user sessions. This similarity measure takes into account the user's interest on visited page. User's interest on visited page is defined by user's relevance on page. Relevance of page is computed by finding TF-IDF value for each page visited by the user. So in order to find user's interest, find frequency of visit to the page by the user in particular session. Interest based similarity measure is used in this work to measure similarity between two weighted session which is given in equation (2).

$$S_{a*b} = \frac{\sum_{i,j=1}^{n} AS_a(I_p)_i . AS_b(I_p)_j}{\sqrt{\sum_{i=1}^{n} AS_a(I_p)_i} * \sqrt{\sum_{j=1}^{n} AS_b(I_p)_j}} \quad (2)$$

This similarity measure is commonly used in case where magnitude of the feature vector is not considered for finding similarity. In our work, we are considering the properties of the vector (i.e interest) for finding similarity. If the similarity between two weighted sessions is 1, then the two weighted session have more similarity. In contrast, if the access pattern of two weighted sessions are orthogonal then the value is 0. Other than this fact, the value may vary from 0 to 1 based on the similarity between two weighted sessions. Once the normal session is converted into weighted session, find similarity between the active user session sequence and subset of session retrieve from the previously logged users' session by using inverted indexed page table. So in order to find similarity, this interest based similarity measure is adopted. The sample similarity between the active user session and each subset of session is given in Table III.

TABLE III
SAMPLE SIMILARITY BETWEEN ACTIVE USER SESSION AND NORMAL SESSIONS

Session ID	**Augmented Session**	**Active User session {0.12,0.9, 0.11}**	**Similarity**
S4	{0.13,0.8,1 ,0.11}		0.25
S6	{0.9,0,8,1}		0.5
S7	{0.9,1,0.6, 0.33}		0.22

iii) *Scalable - KNN approach*

Normally, KNN approach is used to find K closest session to the active user session which suffers from non-scalability problem. To overcome this problem in our recommendation system, indexing technique was adopted. So that only subset of sessions is compared in order to find user interested page. By adopting this technique, the response time of the proposed system is reduced. After computing similarity, the sessions are arranged in descending order based on the calculated similarity value, then the top K sessions are selected and the set of *f* pages which occurs frequently after the sequence similar to the active user session sequence is recommended. For example, consider the session similarity given in table III. After calculating the similarity, order the sessions based on similarity in descending order, So the resultant order would be S6, S4, S7.

TABLE IV
SAMPLE TOP K SESSIONS

Session ID	**List of Visited Pages**
S4	1,3,5,7,8,**11**
S6	1,5,7,8,**11**

Then select top k session on the top of the list, let k = 2, so pick S6, S4 as shown in Table IV. Finally, frequently occurring f page is recommended. Consider f = 1, so the page 11 is recommended to the active user since it occurs frequently after the sequence 5, 7, 8 which is similar to active user session.

IV. VALIDITY MEASURE

Different validity measures are available to check the quality and correctness of the proposed recommendation system. Accuracy is one of the validity measures for evaluating the proposed system. Formula for calculating accuracy is given in equation (3).

$$Accuracy = \frac{Total\ number\ of\ pages\ correctly\ predicted}{Total\ number\ of\ sessions\ in\ testset} * 100 \quad (3)$$

V. EXPERIMENT AND RESULTS

The dataset used in this work is our college web server log data which is of extended log format. This data set contains one month's worth of all HTTP requests from our college web server. The uncompressed content of the dataset contains 3,50,000 numbers of records with timestamp having a one-second resolution. The configuration of the system used for this experiment is 4GB RAM, Intel(R) Core(TM) i3-3217U CPU@1.80 GHZ, running under the Windows-8.1 OS (64-bit).After preprocessing, there is 26.06 % reduction in the size of the dataset. The application was developed using NetBeans IDE and MySQL database. The cleaned records are stored in a relational table with the necessary fields for further process. In order to construct Inverted indexed page table, the constant *w* is important. Multiple run of the same algorithm with different value of sliding window size w with session size greater than six is performed in order to find an optimized *w* value. Cross validation is also performed to find an optimized *w* value. The result is depicted in Table V.

TABLE V
PERFORMANCE OF PROPOSED SYSTEM WITH DIFFERENT VALUES OF *w*

Sliding Window size *w*	**Accuracy**
w = 3	74.47917
w = 4	**81.521736**
w = 5	73.58491

From the above table, it is proven that the constant *w* with 4 is proved to be efficient. So inverted indexed table with *w = 4* is constructed which gives highest accuracy of 81.52 %. Thus our proposed system reduces web users browsing time and recommend page accurately which will be interested by the active user.

VI. CONCLUSION AND FUTURE WORK

The primary objective of the proposed work is to construct an optimized recommendation system which recommends the page based on the active users' exact interest and to reduce the browsing time of the user. Inorder to extract the useful patterns, the weblog file is cleaned and needed fields are extracted. The extracted are stored in database for further process. Users and their list of sessions are identified. Then the inverted indexed page table is constructed which is used to speed up the search process. Inverted indexed page table and previously logged user's session are given as an input to the recommendation system. The normal sessions are reconstructed into weighted session by incorporating user's interest. Similarity between two weighted sessions is calculated by using interest based similarity measure. Then, top k session is selected and set of f pages which occurs frequently is recommended. In our proposed system the user's interest on the visited page is calculated based on frequency. Finally, to increase the prediction accuracy, it might be a very good idea to modify how we are calculating TF-IDF values to include somehow the time spent on a page and not only how many times the page has been requested in a session. The proposed work can also be enhanced by constructing the recommendation system as a distributed one by using parallel programming environment.

ACKNOWLEDGEMENT

The authors wish to thank the Management and Principal of Mepco Schlenk Engineering College, for their support in carrying out this research work.

REFERENCES

[1] D.S.Sisodia, S.Verma and O.P.Vyas,"A Discounted Fuzzy Relational Clustering of Web Users' Using Intuitive Augmented Sessions Dissimilarity Metric",IEEE Journal ,Vol.4,(2016), Page No:2169–3536 .

[2] Abdelghani Guerbas, Omar Addam, Omar Zaarour, Mohamad Nagi, Ahmad Elhajj, Mick Ridley,Reda Alhajj (2013), "Effective web log mining and online navigational pattern prediction",Elsevier journal on Knowledge-Based systems.vol. 49, no. 12, pp. 50–62, Sep. 2013.

[3] H. K. Yogish, G. T. Raju, and T. N. Manjunath, "The descriptive study of knowledge discovery from web usage mining," International Journal of Computer Science Issues (IJCSI), vol. 8, pp. 225-230, Sept.
2011.

[4] S. K. Pani, L. Panigrahy, V. H. Sankar, B. K. Ratha, A. K. Mandal, and S. K. Padhi, "Web usage mining: A survey on pattern extraction from
web logs," International Journal of Instrumentation, Control & Automation (IJICA), vol. 1, 2011.

[5] C. Carmona, S. Ramírez-Gallego, F. Torres, E. Bernal, M. del Jesus, S. García, "Web usage mining to improve the design of an e-commerce website", Elsevier-Expert Syst. Appl. 39(12) (2012) 11243– 11249.

[6] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang Ning Tan" Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data" SIGKDD Explorations-Jan 2000.Volume 1, Issue 2 - page 12.

[7] Xiao J, Zhang Y, Jia X and Li T. Measuring similarity of interests for clustering web-users. In: Proceedings of the 12th

Australasian database conference. IEEE Computer Society, 2001, pp. 107–114.

[8] Sarik Ghazarian, Mohammad Ali Nematbakhsh "Enhancing memory-based collaborative filtering for group recommender systems", Elsevier Journal on Expert Systems with Application,volume:42 2015 ,Page No:3801–3812

[9] Adeniyi D A,Wei Z,Yongquan Y, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method",Elsevier journal on Applied Computing and Informatics (2016) 12, 90–108

[10] Li Chaofeng "Research on Web Session Clustering" Journal of Software, Vol. 4, No. 5, July 2009.

[11] Daqing He, Ayse Goker "Detecting session boundaries from Web user logs" (2000).In Proceedings of the BCS-IRSG 22nd Annual colloquium on Information Retrieval research (pp.57-66).

[12] V.Chitraa, Dr.Antony Selvadoss Thanamani "Web Log Data Cleaning For Enhancing Mining Process" International Journal of Communication and Computer Technologies Volume 01 – No.11, Issue: 03 December 2012 ,2278-9723.

[13] Theint Theint Aye "Web log cleaning for mining of web usage patterns" 3rd International Conference on Computer Research and Development (ICCRD), 2011.

[14] C.-H. Lee, Y.-H. Fu,"Web usage mining based on clustering of browsing features," in: Eighth International Conference on Intelligent Systems Design and Applications, 2008, ISDA'08, 2008.

[15] Abdelghani Guerbas, Omar Addam, Omar Zaarour, Mohamad Nagi, Ahmad Elhajj, Mick Ridley,Reda Alhajj (2013), "Effective web log mining and online navigational pattern prediction",Elsevier journal on Knowledge-Based systems.