A New Web Usage Mining Approach for Website Recommendations Using Concept Hierarchy and Website Graph

T. Vijaya Kumar, H. S. Guruprasad, Bharath Kumar K. M., Irfan Baig, and Kiran Babu S.

***Abstract*—To have a clear and well organized website have become one of the primary objectives of enterprises and organizations. Website administrators may want to know how they can attract visitors, which pages are being accessed most/least frequently, which part of website is most/least popular and need enhancement, etc. Of late, the rapid growth of the use of Internet has made automatic knowledge extraction from server log files a necessity. Analysis of server log data can provide significant and useful information. Information provided can help to find out user intuition. This can improve the effectiveness of the Web sites by adapting the information structure to the users' behavior. Most of the Web Usage Mining techniques use Server log files as raw data to produce the user navigation patterns. Along with the server access log file, we incorporate Website knowledge (i.e., Concept hierarchy and Website Graph) into the web usage mining phases. This incorporation can lead to superior patterns. These patterns can be used to provide set of recommendations for the web site which can be deployed by web site administrator for website enhancement. In this paper, we have considered the server log files of the Website www.enggresources.com for overall study and analysis.**

***Index Terms*—Concept based website graph, concept hierarchy, web mining, web usage mining, website graph.**

I. Introduction

As in conventional Data Mining, the aim of Web Mining is to discover and retrieve useful and interesting patterns from very large web dataset [1]. The World Wide Web (WWW) has become the major source of information in recent years and is growing at humongous rate. All this data on the web can be classified under three different parts, which are Web Structure Mining, Web Content Mining and Web Usage Mining. In this paper we have concentrated on Web Usage Mining (WUM) which can be defined as the application of data mining techniques to web log data in order to discover user access patterns [2], [3].

Web log is a rich source of user's navigation information. Access log recorded at the server side is considered for our study. Any WUM process consists of three vital stages Preprocessing, Pattern Discovery and Pattern Analysis. Preprocessing involves converting user activity information available into data abstractions for pattern discovery [4]. User's navigation behavior recorded in the web server log file contains ambiguity and noise [5]. So to find interesting patterns there is a need to clean these log records and group them into session. Session can be defined as sequence of requests made by a single user using unique single IP address on website for a specified period of time. The usual approach used for session construction is either using only navigation oriented or time oriented heuristics. These two approaches do not effectively capture the actual intention of the user for which he visited the website. After the user identification and session construction, pattern discovery is to be considered as the next step. Pattern discovery involves extraction of the patterns in terms of statistical analysis, association rules, classification, clustering, sequential patterns and dependency modeling. Pattern Analysis is the final stage in WUM following the pattern discovery stage and it involves validation and interpretation of the mined patterns.

Analyzing web logs to extract user's navigation pattern has become necessary for any website administrator to make sure that his site serves the user's needs in a manner preferred by them. Administrator have lots of choices on obtaining users access pattern but, concept based approach can track the actual interest of the users. For example, if a user visits a news website, it would be organized on different concepts like politics, sports, entertainment etc., and then a user interested in sports would only use that concept and later may continue browsing the site or navigate elsewhere. Since the actual interest of the user was in sports, we have to capture it individually for each user and capture how his browsing intent is changing.

The rest of the paper is organized as follows. Section II gives a brief description about the related work. Section III depicts the proposed model. It includes website knowledge, preprocessing, pattern discovery, pattern analysis, and recommendations. The experimental results of the proposed approach are given in Section IV and Section V concludes the discussion.

II. Related Work

Web usage mining refers to the automatic discovery of knowledge from server log files using data mining techniques. Along with the server log file other sources of knowledge such as site content or structure and semantic domain knowledge can be used in Web usage mining [6]. In [7], Natheer Khasawneh et.al have presented new techniques for preprocessing web log data including identifying unique users and sessions by making use of website ontology. In [8],

Manuscript received April 20, 2013; revised July 2, 2013.

The authors are with the Department of Computer Science and Engineering, Bangalore Institute of Technology, Bangalore, Karnataka, India (e-mail: vijaykrte@gmail.com, hs_gurup@yahoo.com, bharathkumarkm@gmail.com, irfanbaig22@gmail.com, kiran.kiranbabu @gmail.com).

Sebastian A. Rios et al. have shown the use of concept-based approach using semantics in web usage mining. In [9], Murat Ali Bayir et al. have proposed a novel framework, called Smart-Miner for web usage mining problem which uses link information for producing accurate user sessions and frequent navigation patterns. In [10], Agarwal et al. introduce problem of mining sequential patterns over databases. In [11], Jaideep Srivastava et al. have proposed a new technique to include conceptual characteristics in WUM recommendation system. The importance of data preprocessing methods and various steps involved in getting the required content effectively is discussed in [12].

III. Proposed Model

Fig. 1 depicts the overall architecture of the proposed model. It involves usual steps of Web Usage Mining such as Data Gathering, Preprocessing, Pattern Discovery, and Pattern Analysis. Our model incorporates website knowledge in web usage mining techniques. Website knowledge is represented via concept based website graph. This is a combination of website graph & concept hierarchy of concerned website. The Website Graph is just like any other graph consisting of vertices and edges, except, the Vertices - represents the Web Pages and Edges - represent the hyperlinks between the web pages. Fig. 2. (a) shows partial website graph.

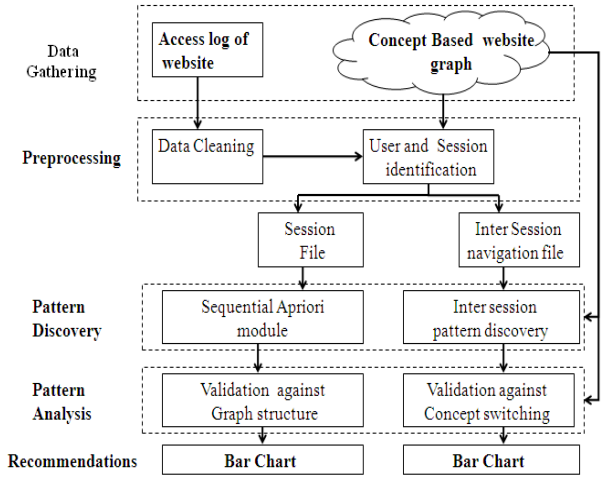

Fig. 1. Overall architecture

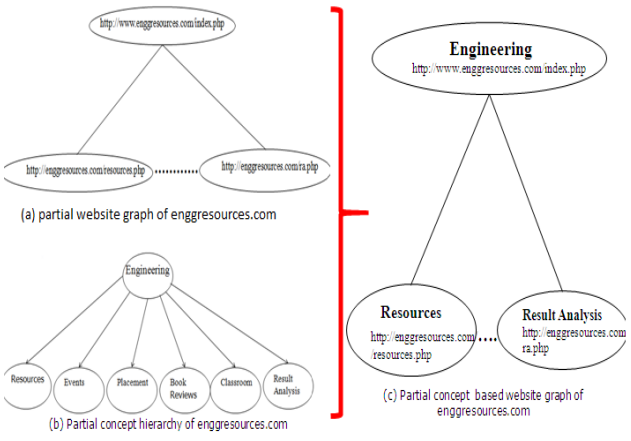

Fig. 2. Construction of concept based website graph

Concept hierarchy represents organization of content of website. Any medium or large websites are usually organized & structured hierarchically to reflect functional characteristics. This hierarchy is collection of domain concepts which are organized using "IS-A" and "HAS-A" relationship. Fig.2. (b) represents the concept hierarchy of enggresources.com. Each concept represents formal abstraction of human thought, his browsing intent. Hence, user's navigination within a session will come under one concept. This information is used as additional constraint to identify session breaking under session reconstruction step. Hence, there is no need for episode identification as extra step.

Concept based website graph (CBWG) is an extension of website graph, where each node will contain information about which concept this webpage belongs to. This mapping of all web pages to corresponding concept is done carefully with help of expertise website knowledge like what functionality this web page provides, what kind of information it depicts etc. CBWG is an ordered pair, $G = (V, E)$ comprising a set V of **vertices** or **nodes,** which represents web pages of website. Every node contains concept as one additional information on which this node is mapped to. Set E of **edges** or **lines**, which are 2-element subsets of V, which contains link information of website graph. Hence, CBWG contains information about conceptual classification of web pages & link information of website. Partial CBWG is shown in Fig. 2. (c).

A. Preprocessing

Preprocessing mainly involves three steps: Data cleaning, User identification, and Session identification. Data cleaning consists of removing superfluous data from log file. Our server log file is in combined log format which is an extension of common log format with two extra fields, referrer & user agent. We have removed error records, requests for images, style sheets, and spider activities. We have retained IP, timestamp, request, and user agent field for further phases.

User identification deals with identifying unique clients to web server. We assume combination of IP & user agent to identify user uniquely. User identification can also be done using client side cookies. But, due to privacy reasons, cookies can be disabled by user, and not every website employ cookies.

Session identification is considered as the next step. A session is a sequence of requests made by a single user with a unique IP address on a particular web domain during a specified period of time. Several methods to identify sessions are discussed in [13]-[15]. We are combining two trivial approaches (time & navigation oriented) with our concept name match approach.

Time oriented approach: The most basic session definition comes with Time Oriented Heuristics which are based on time limitations on total session time or page-stay time. They are divided into two categories with respect to the thresholds they use: In the first one, the duration of a session is limited with a predefined upper bound, which is usually accepted as 30 minutes. In this type, a new page can be appended to the current session if the time difference with the

first page doesn't violate total session duration time. Otherwise, a new session is assumed to start with the new page request.

In the second time-oriented heuristic, the time spent on any page is limited with a threshold. This threshold value is accepted as 10 minutes. If the timestamps of two consecutively accessed pages is greater than the threshold, the current session is terminated after the former page and a new session starts with the latter page.

Navigation oriented approach: Navigation-Oriented approach [16] uses link information of website graph which is present in concept based website graph constructed by using website knowledge. In this approach, it is necessary to have a hyperlink between every two consecutive web page requests.

Let P = [P_1, P_2, . . . , P_k, P_{k+1}, . . . , P_n] be a session containing web pages with respect to their timestamp orders. In this session, for every page P_k, except the initial page $P1$, there must be at least one page Pj in the session which is referring to P_k and has a smaller timestamp than P_k. Topology constraint forces to consider user navigation according to some path in website graph.

Concept-matching approach: This approach considers concepts of web pages from concept based website graph. Adding page P_{N+1} to a session [P_1, P_2, . . . , P_k, P_{k+1}, . . . , P_N] is performed as follows: If concept names of pages P_N & P_{N+1} are same. Then add P_{N+1} to the current session else create a new session and add P_{N+1} to it. i.e., concept switching is taken as one more criteria for breaking session.

Consider {a,b,c,d,e,f} is sequence of requests for web pages by user A identified uniquely by IP & user agent. Similarly, user B requests { u,v,w,x,y,z }. If {a,b,c} belongs to one concept and {d,e,f} to other concept. Then they will be considered as two sessions of user A. similarly for user B, {u,v,w} & { x,y,z} are two sessions if they belong to different concepts. Every session is indexed uniquely to track inter session navigation. Two output files are generated. First, session file, in which one record consists of sequence of web pages requested in a session. Here, it is{{ a,b,c},{d,e,f},{u,v,w},{x,y,z}}. Assume, these sessions are indexed as 0,1,2,3 respectively. Second, inter session navigation file, in which one record consists of session indexes of one user identified uniquely by ip & user agent. Here, this file will have {{0,1}, {2,3}}.

Session construction algorithm(*log_file,CBWG*)

Input: Cleaned log file, concept based website graph

Output: Session file (one line will have one session), inter-session navigation file (one line will have all session indexes of a user)

for each user based on distinct ip & user_agent
for each request of current user
if (time diff b/w cur & prev request < page-stay time threshold **&** time diff b/w cur & first request of session < session time out **&** Link_constraint(prev-request,current-request) **&** Concept_Match(prev-request,current-request))
Add this request to Current Session;
else
Write previous session to session file;
Generate an index for the session and save it ;
Add this request to New Session;
endif
end for
Write All Session indexes of Current user to inter-session Navigation file
end for
end of algorithm.

B. Pattern Discovery

Pattern discovery deals with finding frequent access patterns from sessions discovered in the previous phase. Since conceptual classification of website & considering user navigation within a session lies under a single concept, we identify two kinds of frequent access patterns. They are, frequent navigation patterns within concept, and frequent navigation patterns between concepts (inter session patterns).

To find frequent navigation pattern within concept, we employed sequential Apriori algorithm, which makes use of website topology. In our case, it uses concept based website graph which also holds this topology information of website. We add one extra constraint that uses concept information, to improve efficiency i.e. concept match constraint, which checks before joining, whether the pattern after joining will still belong to same concept i.e. all web pages in that pattern belong to same concept or not. For example, let A = {a,b,c} & B = {d} be two patterns. Before joining pattern A & B, we check {a,b,c,d} belongs to same concept or not. If not then it is pruned now itself before finding out support. Hence, it reduces size of candidate sets further to improve efficiency.

We have considered Apriori algorithm to find frequent navigation patterns between concepts. Here, it is used to find out sequential or inter-session patterns. We also add one extra constraint to improve efficiency that is the concept switching constraint, which expects patterns it is joining, to belong to different concepts. For example, let A={0,1,2} & B={3} be two patterns. Note that 0,1,2,3 are session indexes. Before joining pattern A & B, we check {2} (last item of pattern A) & {3}(first item of pattern B) belongs to different concepts or not. If yes then it is not pruned, else it is pruned. Hence, it reduces candidate sets. This information helps us to capture how user thinking is switching between concepts.

C. Pattern Analysis

Pattern analysis is final stage of web usage mining, which is a descriptive method used to analyze the data such as web usage and customer behaviors. The general summary of the overall user behavior should be shown. It mainly involves validation and interpretation.

Validation is to eliminate the irrelevant rules or patterns and extract the interesting rules or pattern from the output of the pattern discovery process. Interpretation is to represent discovered information in human understanding form.

For navigation patterns within concept, patterns are validated against graph structure using concept based website graph. Interpretation is using bar chart with actual support and confidence for each rule.

For navigation patterns between concepts, patterns are validated for concept switching using concept based website graph. Interpretation is using bar chart with actual support and IP address of users who following that pattern, which can be used in creating user profiles.

Recommendations for website owner can be given in following ways using above results as discussed in [17].

- New content recommendations.
- Website restructuring.
- Page caching/pre-fetching.
- Content enhancements.
- Marketing strategies.

IV. Experimental Evaluation

We have considered the server log of one month (243Mb) from website www.enggresources.com for the experimental procedure, & concept based website graph is prepared as additional input. CBWG consists of 59 distinct web pages which are classified into 9 concepts.

We have used a tool, called Web log filter, to remove unwanted log entries from the log file. Usually, this process removes requests concerning non-analysed resources such as images, multimedia files, and page style files (*.css). For example, requests for graphical page content (*.jpg & *.gif images) and requests for any other file which might be included into a web page and spider activities. By filtering out useless data, we have reduced the log file size to 55Mb. In user identification, IP address and user agent are used. That is, a combination of IP address and user agent is used to identify a unique user. In session construction, we have combined two trivial approaches, Time oriented approach and Navigation oriented approach along with concept name match approach for identifying user sessions. Page stay time threshold and session timeout threshold are set as 10 and 30 minutes respectively. Each web page is assigned with unique index. And, every unique session is also given unique index. Two output files are produced from this step. First, session file, in which one record consists of sequence of web pages requested in a session. Second, inter session navigation file, in which one record consists of session indexes of one user identified uniquely by IP address & user agent. 10217 users and 25814 sessions were discovered from pre-processing.

In pattern discovery, frequent navigation within concept and frequent navigation between concepts are identified. We have employed Apriori algorithm with constraints to improve efficiency in both cases. For frequent navigation pattern within concept, session file from previous stage & CBWG are input, with minimum support threshold as 10% and minimum confidence threshold as 70%. For frequent navigation between concepts, inter session navigation file from previous stage and CBWG are input, with minimum support threshold as 10% and minimum confidence threshold as 70%.

In pattern analysis, patterns are validated with CBWG and interpreted for better understanding with bar charts.

Fig. 3 shows frequent patterns within concept with actual support and actual confidence in terms of bar chart (only few results are shown). Consider a pattern from this figure, /index.php → /resources/resources.php, which has 11.17% support, & we can see that after visiting home page, there is 86.53% confidence of visiting resources page. Hence, browsers first browsing intent is on resources. We can know whether a user is losing interest & aborting his activity from this. We can see from this pattern, most users leaving resources concept without downloading/uploading.

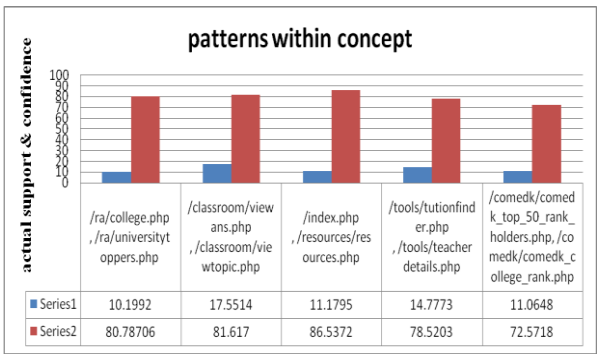

Fig. 3. Frequent patterns within concept

Fig. 4 shows frequent patterns between concepts with actual support and IP addresses of users who follow this pattern (only few results are shown). Consider the pattern from this figure, Placements {/ placements / written.php placements / highscores.php} → resources {/ resources / resources.php}, in which user is switching from concept Resources to Placements to search for written test materials after taking written test. This behavior is observed in 12.5 % users, confidence to switch is 82%, their ip addresses are identified to create user profiles. Recommendation like, direct link to written test materials after taking written test can be provided to user.

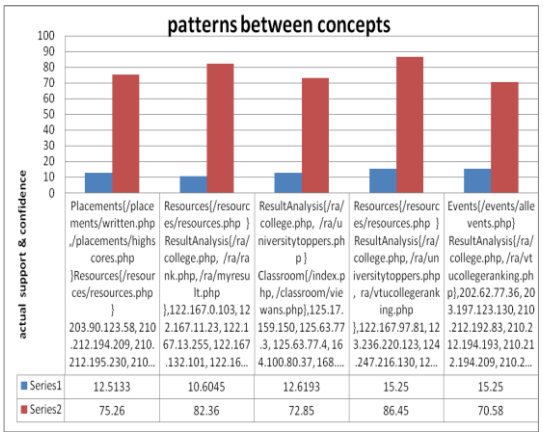

Fig. 4. Frequent patterns between concepts

V. Conclusion and Future Work

In this paper we have introduced a new idea of incorporating available website knowledge for better session construction which would eventually lead to better patterns during pattern discovery. By using concept based approach we can capture the actual intuition of the user which is sole purpose of any mining process. By identifying user's navigation between concepts, we have generated user profiles which will be useful for administrator to predict user behavior for a particular group of users. Recommendation models based only on usage information are inherently incomplete because they neglect domain knowledge. Better predictions can be made by modeling and incorporating

context dependent information: concept hierarchy, link structure and conceptual classification allow us to do so. The results are promising and are indicative of the utility of domain knowledge.

We have created the concept hierarchy from scratch. As a future work, automating this will increase the applicability of our model to a wider class of websites. Without semantic knowledge, recommender systems cannot recommend different types of complex objects based in their underlying properties and attributes. Nor can these systems possess the ability to automatically explain or reason about the user behaviors or user recommendations. The integration of semantic knowledge in terms of website ontology is, in fact, the primary challenge for the next generation of recommendation systems.

REFERENCES

[1] R. Cooley, B. Mobasher, and J. Srivastava, "Web mining: information and pattern discovery on the World Wide Web," *Ninth IEEE International Conference on Tools with Artificial Intelligence*, Newport Beach, CA, USA, 1997, pp. 558-567.

[2] R. Cooley, P. N. Tan, and J. Srivastava, "Discovery of interesting usage patterns from web data," in *WEBKDD*, 1999, pp. 163–82.

[3] J. Srivastava, R. Cooley, M. Deshpande, and P. N. Tan, "Web usage mining: discovery and applications of usage patterns from Web data," *ACM SIGKDD Explorations Newsletter*, vol. 1, 2000, pp. 12-23.

[4] K. E. Amin and N. rouhani, "Web usage Mining:Discovery of the user's navigational patterns using SOM," IEEE, 2009.

[5] R. Cooley, B. Mobasher, and J. Srivastava, "Data preparation for mining World Wide Web browsing patterns," *Journal of Knowledge and Information Systems*, pp. 5-32.

[6] B. Mobasher, "Web Usage Mining in Data Collection and Pre-Processing," *ACM SIGKKD*, Ch. 12, 2007, pp. 450-483.

[7] N. Khasawneh and H. C. Chan, "Active User-Based and Ontology-Based Weblog data preprocessing for Web Usage Mining," presented at IEEE/WIC/ ACM International Conference, 2006.

[8] S. A. Rios and J. D. Velasquez, "Semantic Web Usage Mining by a Concept-based aproach for Off-line Web Site Enhancements," presented at IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008.

[9] M. A. Bayir, I. H. Toroslu, G. Fidan, and A. Cosar, "Smart Miner: A New Framework for Mining Large Scale Web Usage Data," ACM, 2009.

[10] Agarwal and Srikant, "Mining sequential patterns: generalizations and *performance* improvements," in *Proc. the 5th International Conference on Extending Database Technology (EDBT 96)*, France, 1996, pp. 3-17.

[11] B. K. Beemanapalli, J. Srivastava, and S. Sahar, "Incorporating *Concept* Hierarchies into Usage Mining Based Recommendations," *WEBKDD'06*, August 20, 2006, hiladelphia, USA, ACM.

[12] R. M. Suresh and R. Padmajavalli, "An overview of data preprocessing in data and web usage mining," *IEEE*, 2006.

[13] G. T. Raju and P. S. Satyanarayana, "Knowledge Discovery from Web *Usage* Data: Complete Preprocessing Methodology," *International Journal of Computer Science and Network Security*, vol. 8, no.1, January 2008.

[14] E. F. Martinez and V. Karamcheti, "A customizable behavior model for temporal prediction of Web user sequences," in *WEBKDD*, pp. 66-85, 2002.

[15] C. Shahabi and F. B. Kashani, "Efficient and anonymous Web-usage mining for Web personalization," *INFORMS Journal on Computing*, vol. 15, no. 2, 2003, pp. 123-147.

[16] M. Spiliopoulou and L. Faulstich, "WUM - a tool for www ulitization analysis," in *WebDB*, 1998, pp. 184–103.

[17] W. C. Hu, X. L. Zong, C. W. Lee, and J. H. Yeh, "World wide web usage mining systems and technologies," *IEEE*, 2009.

T. Vijaya Kumar is a research scholar at Information Science and Engineering department, BMS College of Engineering, VTU, Belgaum, Karnataka, India. His research interest includes Data Mining, Knowledge Discovery from web usage data, Theoretical Foundation of Computer Science, Artificial Neural Network and Cloud Computing.

H. S. Guruprasad is currently working as professor and Head of Information Science and Engineering department, BMS College of Engineering, Bangalore, India. He has done his Ph.D. in the area of Network Communication. He has more than twenty two years of teaching experience. He has been awarded with Rashtriya Gaurav award and Best Citizen of India award. His research interest includes Multimedia Communications, Sensor Networks, Cloud Computing, Algorithms, Operating Systems, Data Mining and Knowledge discovery from web usage data.

Bharath Kumar K. M. was with the Department of Computer Science and Engineering, Bangalore Institute of Technology Bangalore. His research interest includes Algorithm design, Networking and Web mining. Currently he is working for NetApp, Bangalore, Karnataka, India.

Irfan Baig was with the Department of Computer Science and Engineering, Bangalore Institute of Technology, Bangalore. His research interest includes Data structures, Networking Protocols and data mining. Currently he is working for Wipro, Bangalore, Karnataka, India.

Kiran Babu S. was with the Department of Computer Science and Engineering, Bangalore Institute of Technology, Bangalore. His research interest includes Algorithms, Web Technology and data mining. Currently he is working for Infosys, Bangalore, Karnataka, India.