A Novel Approach for clustering web user sessions using RST

Ms. Jyoti [1], Dr. A. K. Sharma[2], Dr. Amit Goel[3]
Lecturer, Professor and Head, Patent Analyst
justjyoti.verma@gmail.com, ashokkale21@rediffmail.com, goelamit10@yahoo.com,

1, 2 - YMCA Institute of Engineering, Faridabad, (Haryana), India, 3- Noida (Uttar Pradesh), India

Abstract

Web usage mining has assumed importance in learning about web user's behavior and user interactions with the website. It uses data mining techniques to discover non-trivial user behavior patterns. These patterns can then be used to make the predictions of next page to be accessed by the user. Web usage mining consists of the steps like web log preprocessing, pattern discovery and pattern analysis. This paper proposes a novel approach for preprocessing wherein rough set clustering is applied to form the clusters of sessions. These sessions could later on be used to form the knowledge base of rules on the basis of which the next page to be accessed could be prefetched.

I. INTRODUCTION

WWW provides quick and easy access to a tremendous variety of information in remote locations but still users need to wait for the response and if it is delayed beyond 8 sec, they tend to avoid or complain. This user perceived latency is popularly known as "8 sec rule" i.e. if a web page takes more than 8 sec to download then the user switches to other sites [1].

Hence, Reduction of WWW user perceived latency has assumed importance in the wake of the fast development of Internet services and huge amount of network traffic. Web performance can be improved by caching, but the benefit of using it is rather limited owing to filling the cache with documents without any prior knowledge. Web prefetching becomes the solution for this. Prefetching refers to the mechanism of deducing forthcoming page accesses of a client, based on access log information residing at server side or at proxy server [2]. The raw log needs to be preprocessed to find the important sessions that could be used to drill the user information. In this paper, a clustering technique called rough set clustering has been integrated with the preprocessing of the web logs to mine the vital sessions and to formulate them into clusters.

This paper is organized in the following way. The section 2 introduces the framework for prefetching the documents at the proxy side. Preprocessing and rough set theory are discussed in section 3. Section 4 discusses the experimental design using the flowchart. Subsections 4.1 and 4.2 present the pseudo code and the simulated results. Finally section 5 concludes the paper.

II. FRAMEWORK

Internet is a client server architecture where a client sends his request for a resource over the www to a server. The server responds by serving the request. The session involves the exchange of messages and protocols. However, due to exponential increase of www, there are a large number of clients that interact with servers through millions of networks connected with each other leading to a significant increase in the www latency and traffic on the net.

Since a proxy server sits between a Web browser and a web server, it is a potential tool that can be suitably employed to reduce the www latency i.e. it can intercepts all requests to the web server to see if it can fulfill the requests by itself. If not, then only it may forward the request to the web server. In fact, the proxy servers can be employed to achieve two main purposes:

- **Reduce latency:** A Proxy server saves the results of all the requests from various clients for a certain amount of time. For instance, consider a case where both users X and Y access the www through a proxy server. Let us assume that user X requests for a certain web page say Page 1. Sometime later, user Y also requests the same page. Instead of forwarding the request to the web server where page 1 actually resides, which can be a time-consuming operation, the proxy server simply returns this page from its cache where all the downloaded pages are retained before being over written by new arrivals. Since proxy server is often on the same network as the user, this is a much faster operation, thereby reducing the perceived latency to some extent.
- **Filter Unwanted Requests:** Proxy servers can also be used to filter unwanted requests. For example, a company might use a proxy server to

prevent its employees from accessing a specific set of Web sites.

The www latency can be further reduced if the behavior of the user can be predicted and accordingly the predicted pages are prefetched and stored temporarily in the cache of the proxy server. As soon as the user asks for a page, the request can be fulfilled if the requested page is available in the cache. In our work, a prediction engine called Prediction Prefetching Engine (PPE) [3], resides on proxy server as shown in Fig. 1. PPE processes the past references to deduce the probability of future access for the documents accessed so far. The first step of web usage mining is preprocessing the proxy server's log file. A web user may visit several Web sites from time to time and spend arbitrary amount of time between consecutive visits. To deal with the unpredictable nature of Web browsing, web proxy log file should be analyzed

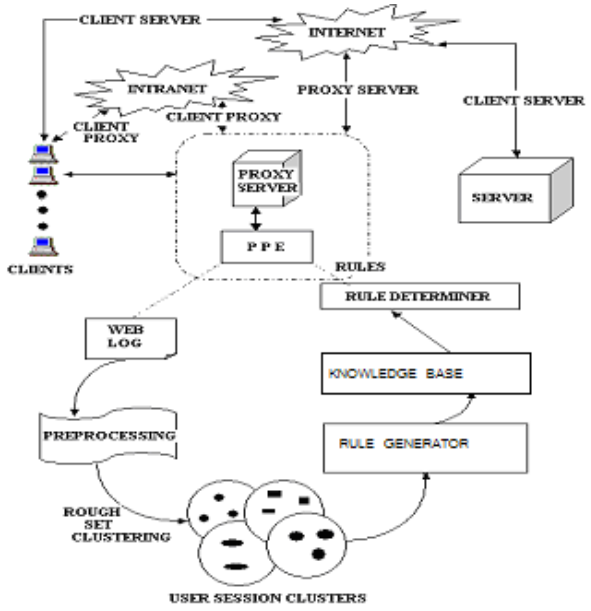

Fig. 1 Prediction Prefetching Engine (PPE), residing on proxy server

The purpose of preprocessing is to transform the raw input data into an appropriate format for subsequent analysis. The input data is stored in the log file that is placed in the proxy server in the above framework. This file must be preprocessed in order to decrease the mining time and also to increase effectiveness in terms of clustering data into sessions. This log file contains all the information of different users e.g. their IP addresses, user id, http request, date and time information, URL requested etc. Log file contain incomplete information, irrelevant data, noise and errors, that need to be filtered out.

After preprocessing of web log, user sessions are clustered on the basis of the maximum time spent. This means that the sessions in which the user has visited the maximum number of pages are the sessions of the highest interest to the user. The proposed work makes use of Rough Set Clustering to cluster the most probable sessions. The advantage of clustering the sessions is that the data mining algorithms can now be applied on the most probable user sessions and hence the rule generator phase will fill the knowledge base with the most appropriate set of rules. The Rule determiner phase of PPE can then fire the rules based on the query that the user enters in the query interface. If the antecedent of the rule matches the query then the consequent of the rule is prefetched and is sent to the proxy cache from where it is sent to the client machine.

This paper focuses on the RST algorithm for clustering the user sessions.

III. RELATED WORK

3.1 Preprocessing

Analyzing the behavior of a Web site's users [4, 8], also known as Web Usage Mining, is a research field which consists of adapting the data mining methods to the records of access log files. Web usage Mining techniques provide knowledge about the behavior of the users in order to extract relationships in the recorded data. Each record in the log file contains the client's IP address, the date and time the request is received, the requested object and some additional information such as protocol of request, size of the object etc. Fig 2 presents a sample of a web access log file. Web browser downloads HTML document on the internet. Several graphics files get downloaded along with other scripts. Generally, user does not make explicit request for such graphics. They get downloaded when web page is requested due to HTML tags. Since web usage mining tracks for the user trails to be followed and since such graphics are not explicitly requested by the user, hence they can be removed. Thus preprocessing of web log consists of sorting, data cleaning [9], user identification and session identification [10].

- **Data Cleaning**: Data is cleaned so as to remove the irrelevant items (such as .gif, .jpeg images).

- **User Identification:** The user's IP address is not sufficient for identifying a user [10]. Many users can be assigned same IP address. Also many users can have access to the same computer. Cookies can be used for better user identification but are not brought into use due to privacy reasons. Moreover, users can block or delete cookies but is estimated that over 90 % of users have their cookies enabled. In such cases, User IDs are brought into use.

- **Session Identification:** A session is the sequence of pages viewed and actions taken by a single user during a defined period of time i.e. 30 minutes [10]**.** Analyzing the web access log and user sessions, user behavior can be understood. By analyzing the user access patterns prediction for the forthcoming page likely to be accessed by the user can be made. This prediction is then used to prefetch that page on to the client cache.

3.2 Rough Set Theory

A lot of previous work has focused on Web data clustering [5, 6, and 7]. Web data clustering is the process of grouping web data into "clusters" so that similar objects are in the same classes and dissimilar objects are in different classes. Its goal is to organize data circulated over the web into groups. The task of clustering is to group users or pages with similar content. Clustering is different from classification, as groups are not predefined. In simple words, it can be said that clustering is used to increase the intra-group interaction and decrease the inter-group interaction. Web data clustering can be categorized into two classes (I) users' sessions-based [7] and (II) link-based [4]. The former uses the web log data and tries to group together a set of users' navigation sessions having similar characteristics. The web log data provides information about activities performed by a user from the moment the user enters a web site to the moment the same user leaves it [8].

Rough Set Clustering (RST) is an approach to aid decision making in the presence of uncertainty [11, 12]. It classifies imprecise, uncertain or incomplete information expressed in terms of data acquired from experience. In RST, a set of all similar objects is called an elementary set, which makes a fundamental atom of knowledge [13]. Any union of elementary sets is called a crisp set and other sets are referred to as rough set (Pawlak 1982). As a result of this definition, each rough set has boundary-line elements. For example, some elements cannot be definitively classified as members of the set or its complement. In other words, when the available knowledge is employed, boundary-line cases cannot be properly classified. Therefore, rough sets can be considered as uncertain or imprecise. Upper and lower approximations are used to identify and utilize the context of each specific object and reveal relationships between objects. The upper approximation includes all objects that possibly belong to the concept while the lower approximation contains all objects that surely belong to the concept.

3.2.1 Nomenclature

A **rough set**, first described by Zdzisław I. Pawlak, is a formal approximation of a crisp set (i.e., conventional set) in terms of a pair of sets which give the *lower* and the *upper* approximation of the original set.

Formally, an information system is a pair $A = (\cup, A)$ where $\cup$ is a non-empty, finite set of objects called the universe and A is a non-empty, finite set of attributes on $\cup$

With every attribute $a \in A$, a set V_a is associated such that $a: \cup \rightarrow V_a$. The set V_a is called the domain or value set of attribute a.

Indiscernibility is core concept of RST and is defined as equivalence between objects. Objects in the information system about which we have the same knowledge form an equivalence relation.

The equivalence relation has the following properties.

If a binary relation $R \subseteq X * X$

- which is reflexive (i.e. an object is in relation with itself xRx),
- symmetric (if xRy then yRx)
- and transitive (if xRy and yRz then xRz)

  is called an equivalence relation.)

Formally any set $B \subseteq A$ there is associated an equivalence relation called B-Indiscernibility relation defined as follows:

$$IND_A(B) = \{(x, x') \in U^2 \mid \forall a \in B\ a(x) = a(x')\}$$

If $(x, x') \in IND_A(B)$, then objects x and x' are indiscernible from each other by attributes from B.

Equivalence relations lead to the universe being divided into equivalence class partition and union of these sets make the universal set.

- Target set is generally supposed by the user.
- Lower approximation is the union of all the equivalence classes which are contained by the target set. The lower approximation is the complete set of objects that can be *positively* (i.e., unambiguously) classified as belonging to target set *X*.
- The *P-upper approximation* is the union of all equivalence classes which have non-empty intersection with the target set. It represents the *negative region*, containing the set of objects that can be definitely ruled out as members of the target set.

IV. EXPERIMENTAL DESIGN

Data preparation includes tasks such as data cleaning,

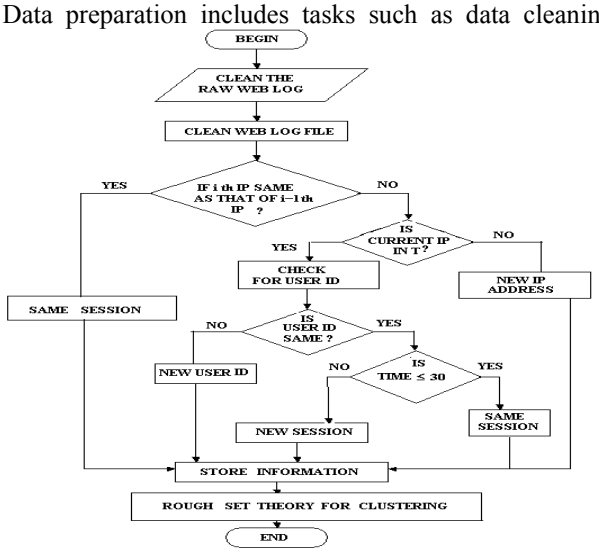

Fig2. Preprocessing and user session clustering process

completeness and correctness, attribute creation and attribute selection. Data conversion must be performed on the initial raw data into a form on which rough set clustering can be applied. The problem is solved in two steps: 1) Preprocessing raw web log to form user sessions. 2) Applying RST over the processed log. Fig. 2 shows the flowchart for the experimental design.

5.1 Pseudo code

A web log is a collection of records of documents accessed by the user. Typical web log record contains following fields: the user's IP address, userID, date and time stamp of the access, HTTP request, the response status, the size of the requested resources, the referrer URL, the user's browser identification. After preprocessing of web log file, user sessions are clustered so as to improve prediction accuracy.

The pseudocode for the proposed work is as:

Clustering_Approximation (WL, ES)
Input WL (Web log table)
Output ES (Equivalence set)
begin
Preprocessing_log (WL, TS)
Input WL (web log table)
Output TS (transaction set table)
begin
j=1;
Remove all the .jpeg and .gif files
TS[j] = WL[i];
for (i=0;i<WL.Length; i++)
if ((WL.ipaddress[i+1]==WL.ipaddress[i]) &&
(WL.time[i+1] - WL.time[i] <=30))
TS[j] =TS[j] $\cup$ WL[i]
endif
endfor
RST (TS, CS) //call Rough Set Clustering algorithm
end
end

RST (TS, CS)
Input TS (transaction set table)
Output CS (clustered set)
repeat
Target_set = { Φ }
Threshold_calculator (TS, threshold)
// call threshold calculator algorithm
for(k=0; k< TS.length; k++)
pages_per_session = $\sum$ TS_k.pages;
if(pages_per_session > threshold)
Target_set= Target_set $\cup$ TS_k
endif
endfor
Matcher(TS, ES) // call Matcher algorithm
Lower_approx = Target_set[p]| $ES_k \subseteq$ Target_set;
Upper_approx = Target_set[p]| $ES_k \cap$ Target_set = Φ
where Target_set has 1,2,…, p elements
forever

Threshold_calculator (TS, threshold)
Input TS (transaction set table);
Output threshold value;
begin
count =0;
for each TS. session in TS
count =count + q
where q = $\sum$ TS.page
threshold = count / $\sum$ TS.session
endfor
end

Matcher (TS, ES)
Input TS (transaction set table)
Output ES (equivalence set)
begin
k=0;
ES = { Φ };
for(i=0; i<=TS.length; i++)
for(j=1; j< TS.length; j++)
if (($j \neq i$) && TS.page[j] == TS.page[i])
$ES_k = ES_k \cup TS_i$;
K=k+1;
endif
endfor
endfor
end

Explanation of the algorithm

Clustering_approximation is the main module which takes as input the web log in the form of table and results

in the equivalence set. To find the desired result, it calls several other algorithms which are briefly explained as below:

- *Preprocessing_log* is the module which takes web log as input and starts with a new empty table transaction set. The algotithm firstly removes all the image files from the web log. The first entry to the TS is the first row of the WL. The algorithm then compares the ip address of the 2nd row of WL with the ip address of the 1st row of WL and it also compares the time interval between the two accesses. If the difference is less than 30 minutes, it implies that both the accesses are of the same session and vice versa. This process is repeated for all the entries of WL and TS is formed accordingly thus separating the different sessions from the WL. Once all the sessions and their entries are marked into the TS, *Preprocessing_log* calls another module which is rough set clustering algorithm.
- *RST (TS, CS)* is the module which takes as input the transaction set and outputs clustered set. It takes Target_set which is empty in the beginning. To fill that target set, threshold value is required. Target_set will contain all those sessions whose number of visited pages is more than this threshold and to find this threshold value it calls another module *Threshold_calculator*. Once the Target_set gets completes, another modukle called *Matcher* is called. Its job is to find the equivalence set. From the equivalence set, lower approximation and upper approximation is found using the formula given in the algorithm.
- *Threshold_calculator(TS, Threshold)* calculates the total number of pages in the various sessions in TS and divides this count by the number of sessions in TS.This value is then used to fill the *Target_set.*
- *Matcher (TS, ES)* finds the equivalence set. In the beginning this set is also empty. It gets filled progressively as the matcher finds all the sessions in TS which contains the same number of visited pages and in the contiguous manner.

5.2 Simulated results

Let after preprocessing the web log, following sessions emerge. Applying the pseudo code, the desired result is shown.

Sessions have been shown on the vertical side and visited pages on the horizontal side.

S1:	P1	P2	P3	P1	P5
S2:	P4	P5			
S3:	P1	P2	P5		
S4:	P1	P2	P2	P3	P4
S5:	P2	P3	P3	P4	P1	P5
S6:	P1	P2	P5		
S7:	P4	P5			
S8:	P1	P2	P3	P1	P5
S9:	P2	P3	P3	P4	P1	P5
S10:	P2	P3	P5		

Now, equivalence classes as per the nomenclature are {{S1, S8}, {S2, S7}, {S3, S6}, {S5, S9}, {S10}}

Threshold > Total pages accessed / n (Sessions)

> 40 / 10 = 4

i.e. Select those sessions which have visited more than 4 pages.

So, TARGET SET {S1, S4, S5, S8, S9}

LOWER APPROXIMATION {S1, S8, S5, S9}

UPPER APPROXIMATION {NULL} because the intersection of the target set with the equivalence classes is empty set so no set qualifies to be in the upper approximation.

Hence, by making use of rough set clustering, we have deduced those user sessions from the web log in which the user spends his quality time. By clustering the important sessions using RST, we have narrowed the web log so that only these sessions could be fed to Rule generator phase of PPE as shown in Fig. 1. The advantage of narrowing the web log is that the complexity of the PPE will be reduced.

V. CONCLUSION

The paper presents a novel approach for finding the user sessions from the web log and then applying rough set clustering to cluster the important sessions based on the maximum pages visited. User sessions are grouped to improve prediction accuracy as data mining techniques are applied on session clusters and not on all sessions and hence the complexity of the PPE is reduced.

VI. REFERENCES

[1] Venkata N. Padmanabhan, Jeffrey C. Mogul ."Using Predictive Prefetching to Improve World Wide Web Latency" , 1997

[2] Y.-H. Wu and A. L. Chen. Prediction of web page accesses by proxy server log. World Wide Web, 5(1):67-88, 2002.

[3] J.Verma, A.K. sharma, Amit goel, " A Framework for Extracting Relevant Web Pages from WWW using web mining", In Proc of International journal of Computer Society and Network security, Seoul, Korea

[4] J. Zhu, Mining Web Site Link Structure for Adaptive Web Site Navigation and Search, PhD thesis, Faculty of Informatics, University of Ulster at Jordanstown, 2003.

[5] P. Baldi, P. Frasconi, and P. Smyth, Modeling the Internet and the Web, Wiley, 2003.

[6] S. Chakrabarti, Mining the Web, Morgan Kaufmann, 2003.

[7] A. Vakali, J. Pokorny, and T. Dalamagas, An Overview of Web Data Clustering Practices, Proceeding of the EDBT Workshop on Cluster Web, Lecture Notes in Computer Science (LNCS) Series, Springer Verlag, Heraklion, Greece, March 2004, pp. 597-606.

[8] Z. Chen, A.Wai-Chee Fu, and F. Chi-Hung Tong, Optimal algorithms for finding user access sessions from very large Web logs. World Wide Web: Internet and Information Systems, 2003, pp. 259-279.

[9] R. Cooley, B. Mobasher, and J. Srivastava, Data preparation for mining World Wide Web browsing patterns Knowledge Information Systems, 1999, pp.5-32.

[10] Z. Chen, A.Wai-Chee Fu, and F. Chi-Hung Tong, Optimal algorithms for finding user access sessions from very large Web logs. World Wide Web: Internet and Information Systems, 2003, pp. 259-279.

[11] Z. Pawlak, Rough sets - theoretical aspects of reasoning about data (Kluwer Academic Publishers, Boston, 1991).

[12] Z. Pawlak, Rough set approach to knowledge-based decision support, European Journal of Operational research 99 (1) (1997) 48-57.

[13] Z. Pawlak, Rough sets, International Journal of Computer and Information Sciences 11 (5) (1982) 341-356.