Ontology Generation from Session Data for Web Personalization

P.Arun
Research Associate, Madurai Kamaraj University, Madurai – 625 021, Tamil Nadu, India.
Email: arun_samy70@rediffmail.com
Dr.K.Iyakutti
Senior Professor, School of Physics, Madurai Kamaraj University, Madurai – 625 021, Tamil Nadu, India
Email: iyakutti@yahoo.co.in

---ABSTRACT---
With an increasing continuous growth of information in WWW it is very difficult for the users to access the interested web pages from the website. Because day by day the information in the web is growing in an increasing manner so without any help system the user may spend more time to get the interested information from the website. To overcome the above problem, in this paper we propose a Model which create a User Interested Page Ontology (UIPO), it will be created by assigning weights and ranking the user interest by count the number of occurrence of each item which was collected from the web logs within a session for all users. The main feature of this model is, it generates UIPO dynamically from that it personalize the interested pages to the web users in their next access The proposed model is very useful for understanding the behavior of the users and also improving the web site design too. The performance of the new model in a session is also discussed in this paper.

Keywords: **Web Usage Mining, Web logs, Ontology, Session, Web Personalization.**

Date of Submission: October 05, 2009 Revised: December 10, 2009 Accepted: January 18, 2010

1. INTRODUCTION.

Data mining is a technique and tools to retrieve the hidden information from the database. The Web mining [1] are the set of techniques of Data mining applied to the web, regarding web mining it is composed of three main areas, Web Content mining is the concept of finding useful information available on-line. The aim of web content mining is to provide an efficient mechanism to help the users to find the information they seek. Web Structure mining is the process of finding the structure of the hyperlinks within the web. Web Usage mining is the process of finding the activities of the users when they are browsing through the web. It is one of the ways of analysis the user's behavior on the website. The goal of the web usage Mining is to personalize[3] the delivery of web content, to improve user navigation, to improve web design, to access the information in fast manner and to improve customer satisfaction.

In our paper we concentrate the web usage mining topic; it is one of the intensive research areas as its potential for personalized services and adaptive web sites. Web Usage mining [1] results mainly depends upon the proper preparation of the data from the web logs. The web log data can be collected from Client side, Server side (or) Proxy servers; here in this paper we collected web logs from server side.

In the recent years the information in the World Wide Web [1] will be increased in an explosive way. From the users point of view the web is a collection of large amount of information and a more portion of time is needed to search and get interested information. To overcome the above problem, in this paper we propose a model for improving the web usage technique in a website and also for the users to collect their interested information in a better way. Here we give more attention for collecting the information on the user sessions from the analysis of HTTP requests made by clients. Usually a user session is a collection of requests made by the user within an interval of time and normally the maximum session time may be of two hours.

Here the Model will create a User Interested Page Ontology (UIPO) [2] [7] [8], it will be created by assigning weights and ranking the user interest by count the number of occurrence of each item which was collected from the web logs in a session for all users, from the UIPO it personalize the interested pages to the web users in their future access. The study of the users' access pattern extracted from the web log files may help the web designer to understand the user behavior, find out the interested object of the website, to find out users ' problems and rearrange the structure and design of the web site based upon it.

In this paper, we propose User Interested Page Ontology (UIPO) model for web personalization, the remainder of this paper is organized as follows: section-2 describes Collecting Web logs from user session for developing the UIPO System, section-3 describes Designing the User Interested page ontology system, section-4 Results and discussion of the UIPO model, section-5 Conclude with a summary and the last section show the list of references with authors biography.

2. COLLECTING WEB LOGS FROM USER SESSION FOR DEVELOPING THE USER INTERESTED PAGE ONTOLOGY (UIPO) SYSTEM .

Here in this paper, we collect the web logs from the reputed website, it has 8 objects and we collect around 10,000 web logs for a period of 60 days. The web log file contains the following entries:

- Date and Time Stamp
- IP Address (Internet Protocol Address)
- URL (Uniform Resource Locator) address of the accessed item.

The following table shows the sample raw web log file.

Table -1
Example of sample raw log file

```
DATE-July 31, 2009, 7:33 am IP-122.164.110.169
FILE NAME-/tender.php
DATE-July 31, 2009, 7:33 am IP-122.164.110.169
FILE NAME-/tender.php
DATE-July 31, 2009, 7:34 am IP-122.164.110.169
FILE NAME-/tender.php
DATE-July 31, 2009, 7:37 am IP-138.198.100.42 FILE
NAME-/index.php
DATE-July 31, 2009, 7:38 am IP-122.174.72.6 FILE
NAME-/tender.php
DATE-July 31, 2009, 7:38 am IP-122.174.72.6 FILE
NAME-/index.php
DATE-July 31, 2009, 7:38 am IP-121.246.12.127 FILE
NAME-/index.php
DATE-July 31, 2009, 7:38 am IP-122.174.72.6 FILE
NAME-/related.php
DATE-July 31, 2009, 7:39 am IP-121.246.12.127 FILE
NAME-/photo_gallery.php
DATE-July 31, 2009, 7:39 am IP-122.174.72.6 FILE
NAME-/epooja.php
DATE-July 31, 2009, 7:40 am IP-121.246.12.127 FILE
NAME-/index.php
DATE-July 31, 2009, 7:40 am IP-122.174.72.6 FILE
NAME-/photo_gallery.php
DATE-July 31, 2009, 7:40 am IP-122.174.72.6 FILE
NAME-/services.php
DATE-July 31, 2009, 7:40 am IP-122.174.72.6 FILE
NAME-/related.php
DATE-July 31, 2009, 7:41 am IP-121.246.12.127 FILE
NAME-/photo_gallery.php
```

From the collected web logs the following steps has to be done to make the raw web logs to a usable one.

Data filtering: It is the process of collecting relevant data for the process.

Data cleaning (Removing Noise and irrelevant data): It is the process of deleting un-useful requests from the log files (i.e.) removing all the data tracked in web logs that are useless for mining process. These requests are usually for images or multimedia files.

User Identification: Identifying the User with their IP Address.

Session Identification: Session Identification is one of the important concepts for this paper, within a time frame if more than one access from the same user (IP Address) it can be belongs to the same session. Different IP address indicates the different user session, if the difference between the timestamp of two successive entries of the same user web log is greater than the prescribed time frame than it will be treated as a new session of that user, otherwise it will be included in the same session. Regarding Session Time it accepts the value in the form of minutes, the minimum value of 1 or higher will be accepted, but the default maximum value is 120.

The next step is to classify the web logs based upon the items which were accessed by each user (IP address). For that first it classifies every item of the website and coded as Numeric based sequence, it was shown in Table-2

Table -2
A numeric based sequence for the items of the website

Code	Links
1	Travel
2	Temple
3	Worship
4	E-Pooja
5	Festivals
6	Tender
7	Gallery
8	Social Services

Once the items of the website was classified, the next step is to collect the click stream data for all users and convert the item of the click stream data to numeric based sequence based upon the table-2. After complete the preprocessing step from the collected web logs the model will take the users who have more web pages accessed in a single session itself. For those users it converts the item of the click stream data into numeric based data and it was shown in table-3.

Table-3
Web Log data for the two users

User Name	Access items
xx	1 2 3 4 2 3 4 3 4 1 3 1 2 4 2 4 1 3 4 2 3 4
yy	1 2 4 2 4 3 1 2 3 4 1 2 3 1 4 3 1

From the numeric based web log data for the above two users, the next step is to count the number of occurrence of each item which was collected from the web logs. Here we count it for the above two users, based upon the count it assign weight and ranking the weblog in the order of weights. The next step is to generate the user interested page ontology based upon the above information.

Ontology [8] [10] is the formal, explicit specificat ion of shared conceptualizations. It is a description of concepts and their relationships that can exist in the domain of interest. It is the easiest way to structure the information through the use of ontology, a link will be created to all the items and it is like a graph of concepts. There are lot of tools are available to create ontology's like onto edit, onto seek, onto maker etc But in general, ontology does not provide the concept of personalization, but in our model, based upon the previous access user information, it ranks the user pages based upon their weights and create page ontology and it focuses the most interested page to the users in their future access. So in our model it creates user interested page ontology (UIPO) for personalization the interested web pages to the users based upon their previous access information (web logs).

3. DESIGNING THE USER INTERESTED PAGE ONTOLOGY SYSTEM.

The next phase is to find out the Interesting Pages for each and every user. It will be finding out based upon the following activities:

In a Session

- How many times the user will access the page (count)
- How the user will access that page either directly or thru any other page
- How much time the user will spend in that page (Entry time – Exit time) and
- What are the activities are done in that page etc...

For each and every user, it counts the number of occurrence of each item which was collected from the web logs, for each item of the user the model will find out the time difference between the entry and exit in each item, the operations which was done in that item and finally the model stores all these information in their corresponding c-file. Based upon the count, time and activity it assign weights for that page and store it in the w-file. It ranking the weights in decreasing order and stored it in the w-file itself, for each and every user their will be a separate c-file and a common w-file where the w-file contains the user-id, item and the corresponding weight. Based upon the ranking it will measure the user's interested web pages for that user and the same procedure will be used to generate the User Interested Page Ontology (UIPO) [2] for all users. The important concept of our model is that the UIPO will be updates dynamically based upon the access of the web pages by the users. It will not be suited for new users because for them there will be no previous access pattern so there is no entry in w-file and there is no c-file for that user, for those cases it will generate a UIPO based upon the user interest which will be collected from the user's profile. Suppose if the user's profile information is not relevant to generate the UIPO than for those users it will generate the overall website ontology .

The final step of the proposed model is to personalize[6] the web pages to the user's (Web Personalization), after completed the above step, it is possible to find out the user's interest, from that information in future if the user will enter into the website, it personalize the Interested Web pages to the users. Suppose during that time if the user was access some more pages which was not accessed during his previous visit, the model will collect those information from the web logs and based upon that it updates the counts and weights in the corresponding users' c-file and the common w-file. From that it measures the ranking and updates that users' UIPO dynamically and the updated version of that UIPO will be personalize to that user during their new visit.

We design a model as shown in figure-3 that generates the UIPO from that it find out the interesting web pages for the users and also it personalize those pages to the users during their next visit.

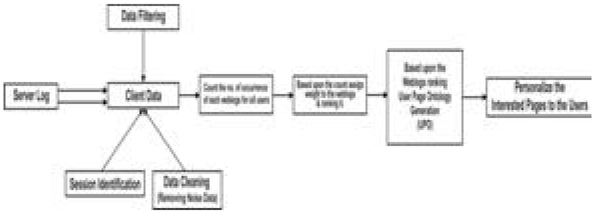
Data Filtering
Server Log
Client Data
Count the no. of occurrence of each weblogs for all users
Based upon the count assign weight to the weblogs & ranking it
Based upon the Weblogs ranking User Page Ontology Generation (UPO)
Personalize the Interested Pages to the Users
Session Identification
Data Cleaning (Removing Noise Data)

Fig. 3. Model generating UIPO for web personalization

The model will be implemented in the C++ Language. Initially

The program collects the transaction from the web logs (objects) for all users but here we take the transaction of above two users and store it in a file.

- Initially the model will assign some value for the support weight.
- The model will assign 60 minutes as the maximum session Time, but during the evaluation & find out the performance time of the proposed model, it set (0.5, 1.0, and 2.0) different time duration for the session time.
- The next step is to count the number of occurrence of the objects for each user which was collected from the web logs and it store in the corresponding users' c-file.
- From the c-file information it count the number of occurrence of the web log, how much time the user was spend in that webpage (time) and what are the activities are done in that webpage, based upon these information the model will assign weight for that item/object.

If the weight is greater than the support weight the model accept that item and store it in the w-file, otherwise it just keep that item/object in that users' c-file itself. So the w-file contains

- User-id, weights and items (which are greater than the support weight value)
- The next step is to sort the items of the w-file according to the order of user-id, weight, from that it shows the user-id, item according to the decreasing order of the weight. So that it is possible to find out that user interested object in the website. The same procedure will be used for the remaining users to find out the interested object of the website. Suppose if the items of the w-file will be sorted based upon the weight from that it is possible to find out the most interested web pages of the users in that website, because here it ranking the WebPages based upon the weight only.

- Based upon the above information it generates the User Interested page Ontology for the corresponding user. The UIPO will be generated for all users based upon their information in the weighted file generated by the previous step.

In future if the user will access the item from the website, than the model will increment the corresponding count in the c-file and if the item count is greater than the support weight than it include that item, weight and user-id into the w-file, for the remaining cases (items) it updates it's weight in the w-file where the item's already available in the w-file itself. This dynamic updating of the user's interest was done in our model.

Ontology [8] is the formal, explicit specification of shared conceptualizations. In general, ontology does not provide the concept of personalization, but in our model it creates User Page ontology for personalization that find out the interested web pages to the users based upon their previous access information.

The User Interested Page Ontology will be generated as follows:

- It creates a link to the most interested web page which was available in w-file and that link will be accessed by that user
- The above process (creating link) is repeated for all the remaining items available in the weighted file(w-file) for that user

From that UIPO, it personalized the most interesting web pages to the users. If we apply the above technique in a website, in future if the user will access the website, it personalizes the interesting pattern to those users. This is the concept of analyzing browsing behavior by collecting basic browsing elements and defining the most interesting pages for each user using user interested page ontology generation (UIPO) for web personalization.

4. RESULTS & DISCUSSION THE NEW MODEL.

There are lot of approaches [6] [7] dealing with web usage mining for the purpose of finding the interesting information (or) automatically discover the user pattern, but in our model the whole process will be divided into three sub process they are

Data preparation which contains the collection of web logs and converts the raw web log data into usable one. In that, counting the items for each user, assign weight and ranking the web log based upon the weight (user interest). Generating User Interested Page Ontology, it will be created from the w-file where the user-id, item will be arranged in the decreasing order of the user-id, weight which was created from the c-file for all users. Once this is successfully completed then the system will provide the positive personalization or recommendation to the user.

The new model was implemented in C++. Here we collect the web logs for 60 days, from that the top two users with most access in a single session data were collected. To find out the performance and comparison of the Existing Model with UIPO model, the following figure-4 shows the users access percentage of the different objects in the website. To find out the user access percentage, from the w-file count the number of occurrence of each item and their corresponding weight (for all users); based upon that information it is possible to find out the access percentage of the object in the website.

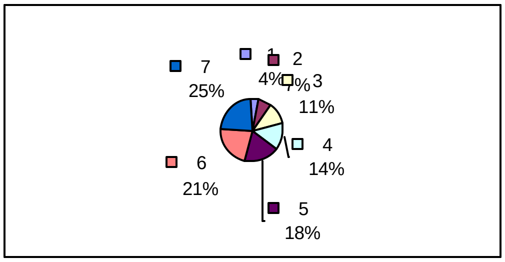

Fig. 4 Performance of the objects in the website

From the above graph it is possible to find out the most interesting objects / most users' access object of the website. From the graph the most user access object is the 7th Object of the website and also it is possible to find out those objects which need to improve more in the website.

To find out the users maximum and minimum number of request per session, the following figure-5 shows that 80% of the sessions have user requests with less than or equal to 20, this means that majority of the session have only small number of requests. The another case is that sessions have the user request greater than 75 that means in total 3% of the session have more than 75 user requests, this will play an important role to study about the browsing behavior of the users.

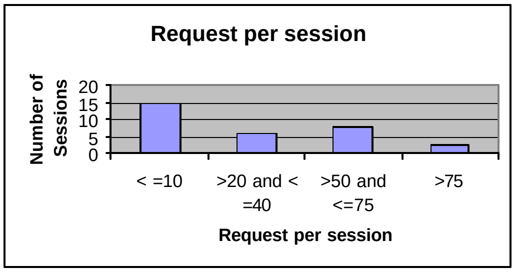

Fig 5. Request per session

To find out the performance of the proposed model with the existing model, we then experimenting the model with predefined period condition and find out the satisfied measures with respect to different period from 0.5 -2.0 hours. The following figure-6 shows the overall performance of the new model for the different session durations from (0.5-2.0 hours) for the two users. In general if the number of personalization increased then the system will work in a fine tune and the overall performance of the system is good and the requirement of the users will be satisfied in an increasing way.

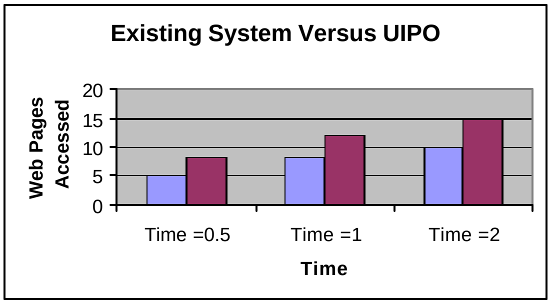

Fig. 6. Existing System versus UIPO

From the above graph we will find out the difference between this method and for the normal case, in our model it recommend the interesting page to the users, from that the users will access the pages in a fast manner. The main purpose of our model is to find out the interesting web pages for the users and also based upon the user's interest it is possible to improve the website design, The main aim of our model is to improve the performance of the access method (i.e.) the personalization process will surely improve the system performance compare to the normal access by the user.

5. CONCLUSION.

The Limitation of our model is, to create User Interested Page Ontology for the new users in the website, because we will create UIPO from the web log information only. For the new users there is no web log data, for those cases it creates the UIPO based upon the user profile or creates the UIPO for the corresponding website without user's interest. In future we will find solution for this problem. Ultimately the aim of our research is according to the discovered pattern to generate recommendations and improve the website Design. The results produced by our research can also provide guidelines for improving the design of web applications too.

REFERENCES

[1] J. Srivastava, R. Cooley, M. Deshpande and P. Tan, Web usage mining: discovery and applications of usage patterns from web data. SIGKDD Explorations, 1(2):12–23, 2000.

[2] S.SenthilKumar and T.V.Geetha, Personalized Ontology for Web Search Personalization, Annual Bangalore Compute Conference, Proceedings of the 1st Bangalore annual Compute conference Bangalore, India ,Year of Publication: 2008 ISBN:978-1-59593-950-0

[3] Kenneth Wai-Ting Leung, Wilfred Ng and Dik Lun Lee, Personalized Concept- Based Clustering of Search Engine Queries. IEEE Transactions on Knowledge and Data Engineering Vol. 20, No11, November 2008.

[4] Gerd Stumme, Andreas Hotho and Bettina Berendt, "Usage Mining for and on the Semantic Web", - Cite seer Institute for Applied Computer Science and Formal Description Methods (AIFB), University of Karlsruhe, D-76128 Karlsruhe, Germany,

[5] Hongyu Zhang, "The Scale-Free Nature of Semantic Web Ontology", School of Software. Tsinghua University. Beijing 100084, China. ACM 978-1-60558-085-2/08/04.

[6] Massimiliano Albanese, Antonio Picariello and Carlo an Lucio Sansone, A web Personalization system based on web usage Mining Techniques, WWW2004, May 17-22,2004, New York, USA, ACM 1-58113-912-8/04/05.

[7] Alexander Maedche and Steffen Staab, Ontology Learning for the Semantic Web, Ontoprise GmbH, Haid-und-Neu-Strasse 7, 76131 Karlsruhe, Germany

[8] K.R. Reshmy and S.K.Srivatasa, Automatic Ontology Generation for Semantic Search System Using Data Mining Techniques, Asian Journal of Information Technology 4(12) 1187-1194, 2005.

[9] Rui G.Pereira dna Mario M.Freire, SWedt: A Semantic Web Editor integrating Ontologies and Semantic Annotations with Resource Description Framwork, AICT/ICIW 2006 0-7895-2522-9/06 @2006 IEEE.

[10] Jeff Z. Pan, A Flexible Ontology Reasoning Architecture for the Semantic Web, IEEE Transactions on knowledge and data engineering, VOL. 19, No. 2. Feb 2007.

[11] Dennis Quan, David R.Karger, How to Make a Semantic Web Browser, ACM 1-58113-844-x/04/05.

Authors Biography

Arun periasamy is a Research Associate of Madurai kamaraj University Madurai India. His research interests are Web mining and Semantic web. Contact him at arun_samy70@rediffmail.com

Iyakutti Kombiah is a Senior Professor of School of Physics of Madurai Kamaraj University, Madurai, India. His research interests are Computational Physics and Software Engineering. Contact him at iyakutti@yahoo.co.in .