International Journal of Computer Applications Technology and Research
Volume 2– Issue 1, 81-85, 2013, ISSN: 2319–8656

Generating User Interesting Page (UIP) Using Supported Weight Value

Nu Yin Kyaw
University of Technology,
Myanmar

Abstract: With an increasing continuous growth of information in WWW, it is very difficult to extract useful and relevant information from the huge amount of information. So, without any help on the system, the user may spend more time to get the interested information from the website. To solve above the problem, we proposed an approach for generating user interesting page (UIP) using weight value on web log data by associating with web usage mining techniques. Web usage mining, a classification of Web mining, is the application of data mining techniques to discover usage patterns from click stream data. This information can be exploited in various ways, such as enhancing the effectiveness of websites or developing directed web marketing campaigns. The goal of our system is to analyze user behaviours by mining enriched web access log data and create a top web page for the user with common needs or interests. This paper also focuses on to provide an overview how to generate frequent access pattern for the users from a Web log database without the use of domain specific ontology.

Keywords: web usage mining, web log data, user session identification, user interesting page, click stream data

1. INTRODUCTION

With the continuous growth and abundance of information on the Internet, the World Wide Web (WWW) becomes a huge repository of information. Nowadays, the Web has become an important medium to communicate ideas, transact business and promote entertainment. The discovery and analysis of useful information from the Web documents is referred to as Web mining [1].

The Web mining [2] are the set of Data mining techniques applied to the web. Web mining can be divided into three categories: web content mining, web structure mining and web usage mining. Web content mining is the process of extracting knowledge from documents and content description.

Web structure mining is the process of obtaining knowledge from the organization of the Web and the links between Web pages. Web usage mining analyzes information about webs pages that were visited which are saved in the log files of Internet servers to discover the previously unknown and potentially interesting patterns useful in the future.

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. Generally, Web Usage Mining consists of three processes: [3] data preprocessing, patterns discovery and patterns analysis. In the data sources of patterns discovery, the results' quality of data preprocessing influences the results of patterns discovery directly. Better data sources can not only discover high quality patterns but also improve the performance of Web Usage Mining. So, data preprocessing is particularly important for the whole Web Usage Mining processes and the key of the Web Usage Mining's quality

At present, since the web becomes the largest unstructured data source available, this condition presents a challenging task for effective design of and access to web pages and need more time to search and get interested information. To overcome the above problem, in this paper, we propose a system for improving the web site design and also for the users to collect their interested information in a better way and then generates personalized web page of their interest dynamically. Today, most web site use ontology-based model for personalization to obtain semantic information. Ontology [6] is a description of concepts and their relationships that can exist in the domain of interest. But in general, ontology does not provide the concept of personalization. So, we want to create a system that generates user interested page (UIP) without the use of the specific domain ontology.

In this paper, we give more attention for identifying clients and collecting the information from the user sessions for 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. In the previous studies, sessions' identification was considered that a user can not be stationed on a page more than 30 minutes. [7] The current study intends to add an improvement in sessions' identification and user's identification with promising algorithms to improve the performance of generating user's interested page. Our proposed system, user interested page (UIP) , will be created by assigning weights and positioning 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 that it personalizes the interested pages to the web users in their next access to the system.

.

2. RELATED WORK

More and more researchers focus on Web Usage Mining recent years. There are lot of approaches dealing with web usage mining for the purpose of finding the interesting information (or) automatically discover the user pattern to improve the purpose of web site design. Pei *et al.* [12] have successfully used the log data from Web logs to discover frequent patterns, they proposed an algorithm called (WAP) Web access pattern tree for efficient mining of access patterns from pieces of logs, Murate *et al.* [13] highlights the importance of analyzing users web log data and extracting their interests of web-watching behaviors and describes a method for clarifying users interests based on the analysis of the site-keyword graph, while Borges *et al.* [14] modeled users' to capture Web navigation patterns. Dr.K.Iyakutti and P.Arun [11] also proposed a web personalized system in order to understand the behavior of the users and also to improve web site design. In this model, user identification is considered under the client IP address only. Session identification is considered using predefined time based

method. In fact, time based method is not appropriate for session identification. They offered the inaccurate performance and results when giving personalized recommendation to users. However, this model has no serious drawbacks. While Spink *et al*. [15] analyzed characteristics of general Web search logs from different perspectives: terms, queries, sessions, and result pages. They showed top users short queries, a small number of search terms were used with high frequencies, few queries were modified, few result pages per query were be visited, and the popularities of query topics.

3. DATA PREPROCESSING

Log files [7] are created by web servers and filled with information about user requests on a particular Web site. These log files are stored in various formats such as Common Log Format (CLF) or Extended Log Format (ELF). Every entry in the log file stores the following fields:

- Client IP address or host name
- Access time
- HTTP request method(GET, POST)
- Path of the resource on the Web server
- Protocol used for transmission
- Status code
- Number of bytes transmitted.
- User agent(browser, operating)
- Referrer

```
95.175.194.33-[27/July/2011]"GET/cuss/home.htmlHTTP/1.1" 200 2553
"http://www.nicelayout.com" "Mozilla/5.0+(compatible;++MSIE+6.0;+Windows+NT+5.1)"

102.175.180.33-[27/July/2011]"GET/cuss/hyperlink.pdf HTTP/1.1" 200 2553
"http://www.nicelayout.com" "Mozilla/5.0+(compatible;++MSIE+6.0;+Windows+NT+5.1)"

97.175.194.33-[27/July/2011]"GET/cusps/announce.html HTTP/1.1" 200 2553
"http://www.nicelayout.com" "Mozilla/5.0+(compatible;++MSIE+6.0;+Windows+NT+5.1)"
```

Figure. 1 A portion of server web log

A portion of the raw web access log files on the server before data cleaning is shown in Figure.1. a successful analysis is based on accurate information and quality of web log data, preprocessing plays an important role.

Data Collecting: we collect the web logs from the commercial website, it has many items. But the web log file we compute contains the following entries:

- Date and Time Stamp
- IP Address (Internet Protocol Address)
- URL address of the access item
- User Agent
- Referrer, etc.

Data Cleaning: The purpose of data cleaning is to remove irrelevant items stored in the log files that may not be useful for analysis purposes. [8],[9]When a user accesses a HTML document, the embedded images and multi-media files are also automatically downloaded and stored in the server log. For example, log entries with file name suffixes such as gif, jpeg, GIF, JPEG, jpg, avi and flv can be removed. This can be done by checking the suffix of the URL name. In addition to this, erroneous files can be removed by checking the status of the request (such as a status of 404 indicates that the requested file was not found at the expected location). A status with value of 200 represents a succeeded request. A status with value different from 200 represents a failed request.

User Identification: A user is defined as the principal using a client to interactively retrieve and render resources or resource manifestations. [11] The Web Usage Mining methods that rely on user cooperation are the easiest ways to deal with this problem. However, it's difficult because of security and privacy. In previous case, user identification is done under IP address heuristic only. IP addresses, alone, are generally not sufficient for user identification. In our paper, we use the subsequent heuristics to identify the user. The following is the algorithm we use to identify individual user in our system. We think that our proposed algorithm will improve the efficiency and the accuracy of user identification.

Proposed Algorithm for User Identification

Input: N entries of web log file

Output: identified User Sets

***Algorithm*:**

While (! last entry of log file)

{

Compare IP address of first log entry with IP address of second log entry.

If (both are same)

Compare the user agent of both entries

If (both agents are same)

Check requested page is linked with the previous access pages.

If (they are linked)

Identify request entries are from the same user.

Else

Assume that they are different users.

Else /* both agents are different */

If (user path traversal is similar as previous one)

Identify request entries are from the same user.

Else

Identify request entries are from the different user.

Else Assume that they are different users. /* IP are different */

} // while loop

Session Identification: A user session can be defined as a set of pages visited by the same user within the duration of one particular visit to a website. Users may have visited the pages for long periods of time. It is necessary to divide the log entries of a user into multiple sessions through a prescribed timeout. The method of portioning into sessions is called as Sessionization or Session Reconstruction. At present, the methods to identify user session include timeout mechanism and maximal forward reference mainly. Time-based session identification can't give accurate user's interesting page.The following is the proposed algorithm that we will use to identify user's session in our system:

Proposed Algorithm for Session Identification

Input: N requests of user set, Traversing Maximum Time, Traversing Minimum Time, 2D array.
Output: Identified Session Sets
Algorithm:
While (! Last entry row of 2D array)
{

Step 1: Calculate the visiting time of a web page of a user.

Step 2: Compare the visiting time with Traversing Maximum Time and Traversing Minimum Time of each web page.

If the visiting time is less than Traversing Maximum Time then assign the weight as 0.

Else if visiting time is between Traversing Maximum Time and Traversing Minimum Time then assign the weight as 1 to 10.

Else if browsing time is greater than Traversing Maximum Time then assign the weight as 100. And if referrer URL is null then weight is assigned as 000.

If the same page is visited by the user again in each user's set then increment the corresponding entry object.

}

To identify user sessions, 2D array is constructed form the user's traversal. Columns are the web pages and rows are users. Visiting time for a particular page is determined by finding the differences between the time fields of two consecutive entries of a same user. Website designers must fix traversing minimum time and traversing maximum time for all web pages as per the contents and loading times. We will compare the visiting time of a user with traversing minimum time and traversing minimum time.
If the value stored is 000 the next entry will be stored as next session in next row. The advantage of assigning weights is that we can observe behaviour of users such as navigation pages, interested pages, and longer duration pages. This information is used to discover interesting patterns of each user for the system.
Data Filtering: It is the process of leaving out less requested resources in the session and retains only the most requested ones. By removing the least requested resources, we can raise performance and accuracy of the system.

4. STAGES FOR GENERATING USER INTERESTING PAGE

Web Log categorizing: It is categorizing of the web logs. For that, it first categories every item of the website and coded as numeric based sequence.

Code	HTML Page Links
1	Shirt
2	Trouser
3	Handbag
4	Cosmetic

Gathering Click Stream Data: It is collecting of the click stream data for each user from each user session.

TABLE I: WEB LOG DATA FOR THE TWO USERS

User	Request Access items
192.168.1.2	123413231342344341
95.102.3.4	1242431234123

Counting Occurrence / Assigning weight & Ranking order: Count the number of occurrence of each item. Based upon the count, the activities that the user makes, it assigns weight and ranking the weblog data in the order of weights**.**
Generating most user interesting page: Find the interesting pages for every user. It generates most_user interesting page for personalization based upon their previous access information (web logs).
Ontology serves as Meta data schemas, providing a controlled vocabulary of concepts, each with explicitly defined meaning. 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 system, based upon the previous access user information, it positions the user pages based upon their weights and create interested web page to the users in their future access. So in our system, it creates user interested page (UIP) for personalization to the users based upon their previous access information (web logs).

5. PROPOSED USER INTERESTING PAGE SYSTEM

Our system intends to find out the Interesting Web Pages for each and every user by analyzing use's click stream data on the web log files. We discover that web page from the identified user's session. From each session, we find top web pages according to the following criteria.

- How many times the user will access the page (count)
- How the user will access that page either directly or through any other page
- How much time the user will spend in that page (Access time – Leave time) and

- What are the activities are done in that page such as (just looking information in the pages, ordering items, registering in the form) 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 system will find out the time difference between the entry and exit in each item, the operations which was done in that item, how many access the web pages etc.

The system stores all these information in user's corresponding record of pre-database. Based upon the count, time and activity it assigns weights for that page and store it in the weight database. It positions the weights in decreasing order and stored it in the weight database. For each and every user, their will be a separate record in pre-database and a weight database where the weight database contains the user-id, item and the corresponding weight. Based upon the position, it will measure the users' most interested web page in the web site and the same process will be used to generate the User Interested Page (UIP) for all users.

The important concept of our system is that our UIP will be updated dynamically based upon the access of the web pages by the users even when they change their desires. The final step of the proposed model is to generate personalized web pages to the user after completing the above steps. We can also find the user's interested web page for next times when the user will enter into the website. Suppose during that time if the user will access some more pages which were 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' corresponding record of pre-database and the weight database. From that it measures the positioning and updates that users' UIP dynamically and the updated version of that UIP will be personalized to that user during their next visit.

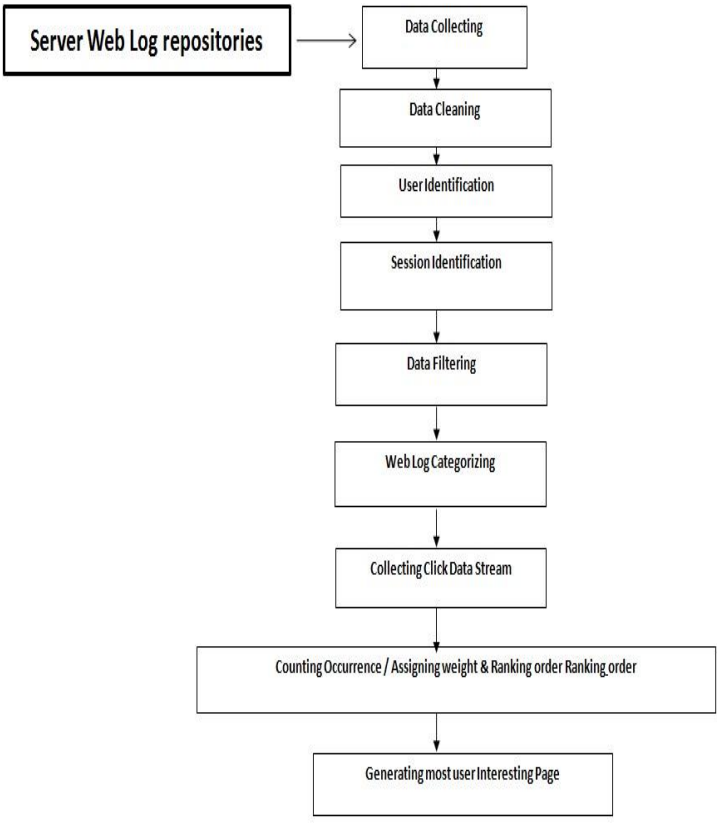

Figure. 2. A proposed system that generating UIP for web personalization

We design our system as shown above figure-2 that generates the UIP 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. And this system will be implemented in the Java Language.

Our system will assign some value for the supported weight on the corresponding items. If the weight of the access web page is greater than the support weight assigned by the system, the system accept that item and store it in the weight database, otherwise it just keep that object in that users' corresponding record of pre-database.

In future, if the user will access the item from the website, then the system will increment the corresponding count in the record of pre-database and if the item's count is greater than the support weight, then it updates its weight in the weight where the item's already available in the weight database itself. This dynamic updating of the user's interest was done in our system.

Ontology is the formal, explicit specification of shared conceptualizations. In general, ontology does not provide the concept of personalization, but in our system it creates User Interested Page for personalization that find out the interested web pages to the users based upon their previous access information. The User Interested Page will be generated as follows:

- It creates a link to the most interested web page which was available in weight database 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 database for that user.

From that UIP, it personalizes the most interesting web pages to the users. If we apply the above methods in a website, in future, the user will access the website, it personalizes the interesting pattern to those users without wasting their time with fine accuracy. 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 (UIP) generation system for web personalization.

6. PROSPECTS AND DISCUSSION OF OUR PROPOSED SYSTEM

Our system will guide the user to find their interested information they want in accurate and fast manner without browsing the whole web site. There are several approaches dealing with web usage mining for the purpose of finding the interesting information (or) automatically discover the user pattern. But in our system, the whole process will be divided into nine sub processes they are: 1.Data Collecting 2.Data Filtering 3.Data Cleaning 4.User Identification 5.Session Identification 6.Web Log Categorizing 7.Collecting Click Stream Data 8.Counting Occurrence/ Assigning weight & Ranking order 9. Generating most user interesting page. Our proposed system can be simple. This system may generate personalized user most interested web page without using privacy information of users. It also help the web designer which page category are getting top for user's interest and which one is less for users and also improve the web site design too.Once this is successfully completed then the system will provide the positive personalization or recommendation to the user. This new system will be implemented in Java Language. And we will collect log files from the commercial web server.

By implementing our UIP system with above promising algorithms for user and session identification, we believe that

our system will be difference than existing cases when generating user's interested information. Our system will also perform the actions with good accuracy, performance and fast manner in accessing the desired web pages.
The main purpose of our system is to find out the interesting web pages for the users based upon the user's interest. It may also be possible to improve the website design. The main aim of our system is to improve the performance of the access method for the website, (i.e.) the personalization process will surely improve the system performance when compared to the normal access by the user.

7. CONCLUSION

At the beginning we present on data preprocessing which has been performed on the log files. Here we presented the methods that we proposed for session and user identification with algorithms. Having the data preprocessing step done, we can then go to other important steps for information mining, the one of effectively extracting useful information from the raw web log data. Mining web important information from web site pages is an important task as it helps web site designers to improve the design of the site. It gives better satisfaction for the final user. By mining most user's interested web pages from web logs; the web site designer can discover the bad web page in the web site and can change the design. Our system presents different ways of solving this problem with better performance. The novelty brought by this work will be implemented by the Java application with a friendly graphical user interface. When implementation is completed, our system can be helpful for site developer in order to arrange the pages and so bring customer satisfaction and increase sells. This also will help web site designers and developers to improve the initial design created and so attract more visitors by the user friendly interface developed. So, the web site designers can determine the web pages that are not correct located and bring them to the right position. But our proposed system will depend on professional web designer to assign weight on every web page of the web site to generate interesting web pages. When our current proposed system is finished completely, our practical experimental results will suggest the significance of the proposed approach.

8. REFERENCES

[1] C.P. SUMATHI, R. PADMAJA VALLI, T. SANTHANAM,"An overview of preprocessing of web log files for web usage mining", Journal of Theoretical and Applied Information Technology, 15th December 2011.

[2] 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.

[3] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan. "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", SIGKDD Explorations, 2000, Vol. 1(2):1-12.

[4] S.SenthilKum ar 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.

[5] Alexander Maedche and Steffen Staab," Ontology Learning for the Semantic Web", Ontoprise GmbH, Haidund-Neu-Strasse 7, 76131 Karlsruhe, Germany.

[6] 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.

[7] D.Claudia Elena**,** "Association and Sequence Mining in Web Usage", Annals of "Dunarea de Jos" University of Galati Fascicle I. Economics and Applied Informatics,1 June 2011.

[8] R.Cooley, Bamshad Mobasherand Jaideep Srivastava, "DataPreparation for Mining World Wide Web Browsing Patterns." Knowledge and Information Systems, 1(1), 1999, 5-32.

[9] R.Cooley, B. Mobasher and J. Srivatsava, "Web mining: Information and pattern discovery on the World Wide Web." 9th IEEE Inernational Conference on Tools with Artificial Intelligence. CA, 1997, 558-567.

[10] Li Chaofeng," Research and Development of Data Preprocessing in Web Usage Mining", Journal of Wuhan 430074, P.R. China.

[11] P.Arun, K.Iyakutti," Ontology Generation from Session Data for Web Personalization", Int. J. of Advanced Networking and Application 241 Volume: 01, Issue: 04, Pages: 241-245 (2010).

[12] J. Pei, J. Han, B. Mortazavi-Asl, H. Zhu, "Mining access patterns the efficiently from web logs" in PADKK '00: Proceedings of the 4 Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications. London, UK: Springer- Verlag, pp. 396-407, 2000.

[13] T. Murata and K. Saito, "Extracting Users Interests from Web Log Data", Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, Hong Kong, China, pp.343-346, 2006.

[14] J. Borges, M. Levene, "A Fine Grained Heuristic to Capture Web Navigation Patterns," ACM SIGKDD Explorations, Vol.2, No.1, pp.40-50, 2000.

[15] A. Spink and B.J. Jansen, "Web search: Public searching on the Web", Dordrecht: Kluwer Academic, 2004.