Adaptive Personalization

Pooja Mehta
Department of Information Technology
Sabar Institute of Technology For Girls
Gujarat,India
poojamehta810@gmail.com

***Abstract*— Web mining is the application of the data mining which is useful to extract the knowledge. Web mining has been explored to different techniques have been proposed for the variety of the application. Web usage mining is used to mining the data from the web server log files. Web Personalization is one of the areas of the Web usage mining that can be defined as delivery of content tailored to a particular user or as personalization requires implicitly or explicitly collecting visitor information and leveraging that knowledge in your content delivery framework to manipulate what information you present to your users and how you present it. Web personalization allows online merchants to customize web content to serve the needs of individual customers. Preprocessing, Page classification and site recommendation are three main steps in this. Our research has been evaluated on academic website. Experiments show that the approach is efficient and practical for adaptive web site.**

Index Terms— web mining, web usage mining, adaptive web personalization

I. INTRODUCTION

Data Mining (the analysis step of the Knowledge Discovery in Databases process, or KDD), a relatively young and interdisciplinary field of computer science, is the process of discovering new patterns from large data sets involving methods from statistics and artificial intelligence but also database management. Web mining is the application of data mining techniques to extract knowledge from Web data including web documents, hyperlinks between documents, usage logs of web sites, etc [1].

Two different approaches were taken in initially defining Web mining. First was a 'process- centric view', which defined Web mining as a sequence of tasks [2]. Second was a 'data- centric view', which defined Web mining in terms of the types of Web data that was being used in the mining process [3].

We decompose the web mining to following sub tasks [4]:

- Resource Finding: the task of retrieving indented Web documents.
- Information Selection and Pre-processing: automatically selecting and pre-processing specific information from retrieved web resources.
- Generalization: automatically discovers general patterns at individual web sites as across multiple sites.
- Analysis: validation and /or interpretation of the mined patterns.

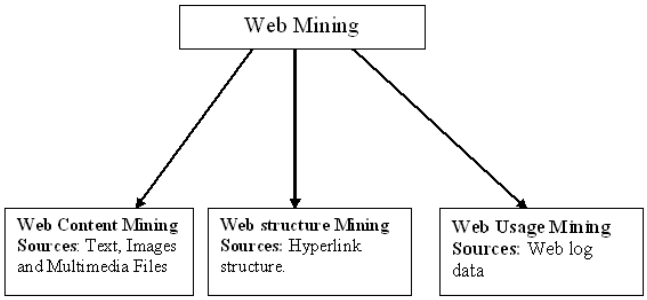

Fig. 1. Types of Web Mining

The above fig (1) shows the types and sources of Web mining. Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables [5]. Research in web content mining encompasses resource discovery from the web, document categorization and clustering, and information extraction from web pages [6]. Web structure mining studies the web's hyperlink structure. It usually involves analysis of the in-links and out-links of a web page, and it has been used for search engine result ranking. [6]. Web Structure Mining can be regarded as the process of discovering structure information from the Web. This type of mining can be performed either at the (intra-page) document level or at the (inter-page) hyperlink level [5]. Web structure mining is the process of inferring knowledge from the World Wide Web organization and links between references and referents in the Web [7].

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web based applications. It also called as Web log mining. Some of the typical usage data collected at a Web site includes IP addresses, page references, and access time of the users. [5]

A. *Areas of Web Usage Mining:*

- System Improvement
- Site Modification
- Business Intelligent
- Usage Characterization
- Personalization

The remainder of the paper is structured as follows. First section reviews phases of Web usage mining. In other section, describes Web Personalization and its processes. Then we research on steps to implement the adaptive personalization.

II. RELATED WORK

Web personalization is a strategy, a marketing tool, and an art. The objective of a Web personalization system is to "provide users with the information they want or need, without expecting from them to ask for it explicitly" [8]. Personalization requires implicitly or explicitly collecting visitor information and leveraging that knowledge in your content delivery framework to manipulate what information you present to your users and how you present it. A personalization mechanism is based on explicit preference declarations by the user and on an iterative process of monitoring the user navigation, collecting its requests of ontological objects and storing them in its profile in order to deliver personalized content [9].

A. *Phases of Web Usage Mining*

The Web usage mining divides in to three distinct phases [1].

- Preprocessing – Data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user. There are types of preprocessing techniques related to web mining category. That are Usage Pre-Processing, Content Pre-Processing, Structure Pre-Processing.
- Pattern Discovery – Web Usage mining can be used to uncover patterns in server logs but is often carried out only on samples of data. The mining process will be ineffective if the samples are not a good representation of the larger body of data. the pattern discovery methods like Statistical Analysis, Association rules, Clustering, Classification, Sequence Patterns, Depending Modeling.
- Pattern Analysis – This is the final step in the Web Usage Mining process. After the preprocessing and pattern discovery, the obtained usage patterns are analyzed to filter uninteresting information and extract the useful information.

B. *Web Data*

Web data are those that can be collected and used in the context of Web personalization. [8][10]

- Content
- Structure
- Usage
- Usage Profile

C. *Web Logs*

Log files are files that contain list the action have been occurred. These log files reside in the web server .The Web Server store all the files necessary to display the web pages on the user computer.

80.84.1.24[31/Jul/2011:17:41:58+0530]"GET/index.php?file=contain&id=127 HTTP/1.0" 200 1380 "http://www.skpharmacycollege.org/index.php?file=contain& id=146" "Opera/9.50 (J2ME/MIDP; Opera Mini/5.1.24700/25.683; U; en)"

Above is the example of web log file which reflect the below information [11].

- IP address: An IP address is 32-bit host addresses defined by the Internet Protocol. One IP address is usually defined for one domain.E.g.80.84.1.24
- Authuser: Username and password if the server requires user authentication.
- Entering and exiting date and time : 31/Jul/2011:17:41:58 +0530
- Modes of request: GET, POST or HEAD method of CGI (Common Gateway Interface)
- Status: HTTP status code returned to the client, e.g., 200 is "ok".
- Bytes: The content length of the document transferred. E.g 13807
- Remote log and agent log: E.g Opera/9.50 (J2ME/MIDP; Opera Mini/5.1.24700/25.683; U; en)"
- RequestedURL: http://www.skpharmacycollege.org/index.php?file=contain&id=146

III. WEB PERSONALIZATION PROCESS

The Web personalization process divides in to four distinct phases [5].

A. *Phases of Web Personalization*

- Collection of Web data: In this, implicit data includes past activities/click streams as recorded in Web server logs and/or via cookies or session modules. Explicit data usually comes from registration forms and rating questionnaires.

- Preprocessing of Web data: In this, Data is frequently pre-processed to put it into a format that is compatible with the analysis technique to be used in the next step. Preprocessing may include cleaning data of inconsistencies, filtering out irrelevant information according to the goal of analysis. Most importantly, unique sessions need to be identified from the different requests, based on a heuristic, such as requests originating from an identical IP address within a given time period.

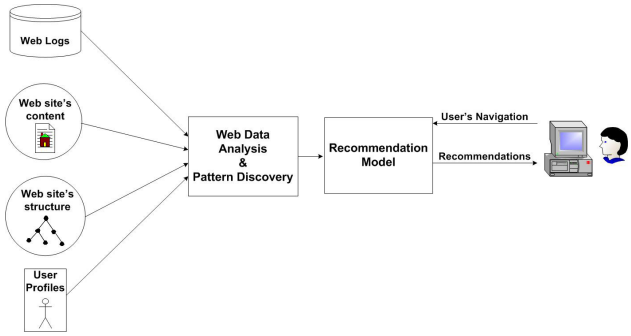

Fig.2. Phases of Web Personalization

- Analysis of Web Data: Also known as Web Usage mining, this step applies machine learning or data mining techniques to discover interesting usage pattern and statistical correlation between web pages and user groups. This step frequently results in automatic user profiling, and is typically applied offline, so that it does not add a burden on the web server.
- Decision making/Final Recommendation: It makes use of the results of the previous analysis step to deliver recommendations to the user. It involves generating dynamic Web content on the fly, such as adding hyperlinks to the last web page requested by the user. This can be accomplished using a variety of Web technology options such as CGI programming.

B. Steps of Adaptive Web Personalization

To build adaptive Web sites by evolving site structure to facilitate user access. To build adaptive web site, we can also use the preprocessing, page classification and site reorganization [12].In preprocessing, pages on a Web site are processed to create an internal representation of the site. Page access information of its users is extracted from the Web server log. In page classification, the Web pages on the site are classified into two categories, index pages and content pages, based on the page access information. After the pages are classified, in site reorganization, the Web site is examined to find better ways to organize and arrange the pages on the site [12].

a) Preprocessing:

In this, pages on a Web site are processed to create an internal representation of the site. Page access information of its users is extracted from the Web server log. To make adaptive web site based on the site reorganization we follow three preprocessing approach.

- Web site preprocessing
- Server log preprocessing
- Access Information collection

b) Page Classification

In this, we divide us pages in to two categories. I) Index page and II) content page. The Pages which is used by the user for the navigation of the system is called the index page.

Below is the algorithm for the page classification:

(1)Find the mean length λ, where λ = 1/ (mean reference length of all pages)

(2) t= -ln(l-y) /λ□,where t=reference length

(3) For each page p on the web site

(4) If p is not HTML or

(5)p's session count >count_threshold , where count_threshold = content page

(6) then p is content page

(7) Else check p's number of link >link_threshold, where link_threshod = index page

(8) then p is index page

(9) else if p's reference length < t

(10) then p is index page

(11) else if p is content page

The Content page is a page containing information in which the user interested. The Session count of a page is ration of the number of time it is the last page of session to the number of session.

c) Site Recommendation

In this, frequently accessed pages are put higher up in the site structure. On other side, infrequently accessed pages are placed lower in the site structure. In site recommendation, the pages are examined sequentially starting from the home page. Below is algorithm for the site recommendation.

(1) Initialize a queue Q

(2) Put children of the home page in Q

(3) Mark the home page

(4) While Q not empty

(5):= current_page = pop (Q)

(6) Mark current_page

(7) For each parent p of current_page

(8) Push children (maybe merged) of current_page into Q if they are not marked.

The goal of the above algorithm is to set the recommendation or preference system as per the user interested. The basic idea of recommendation system is to remove the intermediate index page as a user go through.

IV. EXPERIMENTAL EVOLUTION

we evaluate our approach with an experimental on the web site http://www.skpharmacy.ac.in.About 2 months log file

from July 2011 to August 2011.This web server log is about 14 MB size.

There are various steps for adaptive web personalization. We can measure the parameters in the three different steps like Preprocessing, page classification and site recommendation.

A. *Preprocessing*

In this phase, it cleans the data means it removes all duplicate and redundant information from the web log file. We will measure the parameter like support and confidence for the different section from the web site.

We count the support of all the data in the web log file. The department, photos etc are main pages of the college web sites.

B. *Page Classification*

In this phase we will measure the accuracy and efficiency of the link_threshold, y, count_threshold. We divide the pages in to either index page or content page. In this phase we also count the session, Session per IP and Reference Length.

C. *Site Recommnedation*

In this, system gives the preference or recommendation system to the user. For site recommendation, we will measure the accuracy, performance, efficiency of the F- (frequent page), I-(index page) and C-(content page).

V. CONCLUSION

Web Personalization technique for web usage mining purpose has been discussed since last many year to improve making web sites adaptive. We summaries that many methods of the adaptive web sites are proposed but most of the techniques are not suitable for adaptive web site because they give some issue like less response time, accuracy, frequency, user access patters etc. There is scope for further improvement in proposed methods and algorithm for site recommendation. Future work will be focus on the implementation to get the accuracy and implementation of the artificial and e-commerce data.

REFERENCES

[1] J. Srivastva, P. Desikan and V. Kumar, Web mining – Concepts, Application and Research direction

[2] O. Etzioni. The World-Wide Web, Quagmire or Gold Mine? Communications of the ACM, 39(11):65–68, 1996.

[3] R. Cooley, J. Srivastava, and B. Mobasher. "Web mining: Information and pattern discovery on the World Wide Web". In Proceedings of the 9th *IEEE International Conference on Tools with Artificial Intelligence(ICTAI'97)*, 1997.

[4] R. Kosala, H. Blockeel , Web Mining Research: A Survey, *SSIGKDD Explorations,ACM SIGKDD*,July 2000.

[5] A. Jebaraj Ratnakumar, "An Implementation of Web Personalization Using Web Mining Techniques",Journal of Theoretical and applied information technology.,2005

[6] W. Bin, L. Zhijing ,**"**Web Mining Research". Proceedings of the fifth *International Conference on Intelligence and Multimedia Applications (ICCIMA'03),*2003.

[7] Q. Han, X. Gao, W. Wu ,"Study on Web Mining Algorithm Based on Usage Mining",2010.

[8] M. Eirinaki and M. Vazirgiannis Athens University of Economics and Business, "Web Mining for Web personalization."

[9] D. Antoniou, M. Paschou, E. Sourla, A. Tsakalidis,"A Semantic Web Personalizing Technique The case of bursts in web visits", *IEEE Fourth International Conference on Semantic Computing*,2010

[10] Chen L, Sycara K. "A Personal Agent for Browsing and Searching". 2nd International Conference on Autonomous Agents, Minneapolis/St. Paul, pp132-139, 1998.

[11] Dr. G. K. Gupta, "Introduction to Data Mining with Case Studies" , *PHI* Publication

[12] Fu Y., Yi M. S., Creado M., "Reorganizing websites based on user access patterns", International Journal of Intelligent Systems in Accounting, Finance and Management 11,pp 39-53, 2002.