A SURVEY ON WEB PERSONALIZATION IN WEB MINING

R. Gobinath[1], M. Hemalatha[2]

ABSTRACT

Internet plays a major role in information retrieval in this modern day. Effective handling of web technology coupled with various techniques provides a better way for users to surf websites. Web mining techniques, the most likely used application of data mining in the web personalization process for user behavioral analysis. The behavioral patterns gathered from the log files may be used for the improvement of the websites and also helpful for the website owners to improve business. The necessary information gathered from the log files provide actual behavioral patterns followed by historical users. This paper focusses on the steps involved before the analysis of gathering navigational patterns and after pattern analyzing.

Key words : World Wide Web, Web Log Files, Web usage mining, Web Personalization.

I. INTRODUCTION

Many organizations, business institutions and companies are very keen on identifying users' behaviours for providing excellent services to the public by improving their product. The web site is one of the sources for collecting user navigation patterns which are recorded as navigational history in server which is known as web log files. These access log files from the server which are useful in identifying the behavioral patterns for web personalization. The e-commerce websites, e-learning websites and e-banking websites are very interested in gathering such users' behavior information for improvement.

The visitor-customers face a complexity in using web sites because of information overload. Website usage can be increased by providing a user-friendly environment to the customers. The user-friendly environment can be achieved in the construction of web pages by personalizing the web according to the behavioral patterns of historic users. The pattern history gathered from the previous users serves as a mode for understanding the interest about the web page. The customer's needs can also be analyzed in web personalization and can be used in e-business to provide good service to the customer.

Web personalization also covers areas like recommender system, customization and adaptive web sites. Many existing commercial personalization systems involve manual work and consume more time. The numbers of personalizing web pages are increasing day by day to solve this problem [3]. Telecommunication and many fields of entertainment are also started to use personalization technologies broadly. The web personalization technology can be expanded by automating the web based

[1]Research Scholar, Department of Computer Science, Karpagam University, Coimbatore.
E-mail : iamgobinathmca@gmail.com

[2]Professor, Department of Computer Science, Karpagam University, Coimbatore.
E-mail : hema.bioinfo@gmail.com

services and it should be adapted by the user. Machine learning methods [7] have adapted personalization techniques and have successful records. Data mining can provide a complete solution to the adaptation task.

Knowledge Discovery in Data (KDD) has been used to analyze data collected on the Web and extract the necessary information. The process of extracting necessary information from huge database can be done with web mining application by implementing data mining process and the analysis of such data can be done using web usage mining [12],[13]. This paper focusses on web usage mining. The usage mining is widely recognized for its better solution and to provide ideas for web site personalization. This survey is helpful for the researchers of usage mining to benefit some knowledge about web personalization process. Section II explains Web mining, section III explains about the personalization, section IV explains about the challenges in web mining personalization, section V explains uses of web personalization, and section VI deals with the conclusion.

II. Web Mining

The necessary information extracted from the web is known as Web Mining i.e Fig 1. The knowledge framework is extracted from the web and given a proper representation. We can extract various forms of information from the web by using web mining application. Web mining application for extraction of information was evolved to satisfy the previous difficulties of extraction of information from web. The evaluation in the web mining is advanced in extracting the information of multimedia [14]. The discovery of patterns from the web can also be performed by web mining application.

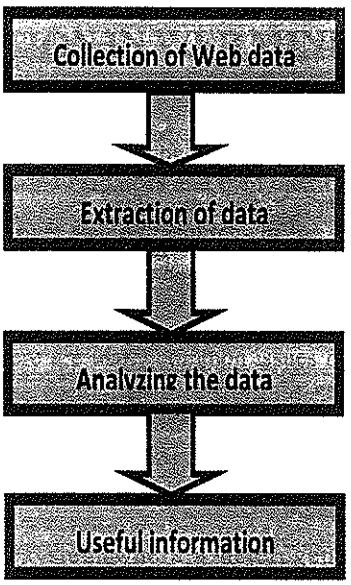

Figure 1 : Web mining steps

Web mining steps:

- Collection : Retrieve the content from the web
- Extraction : Extract data from formats
- Analysis : Tokenize, Rate, Classify, Cluster
- Knowledge : Necessary information

There are three major areas in Web Mining [33]. They are:

a) Web Content Mining

b) Web Structure Mining

c) Web Usage Mining

A. Web Content Mining:

Web content mining is related to but different from text mining and data mining. Web content mining uses many data mining techniques and has a strong relation with the data mining. Many of the web contents are texts and web

content mining is also related to text mining. Data mining deals with the structured data which differs from the web content mining, because web content mining deals with unstructured web data. Text mining deals with unstructured texts but web content mining deals with the semi-structured web data. Web content mining approaches should have creative data mining and text mining applications.

Web content mining contents are

- Web page
- Search page
- Result page

Web Page :

A Web page carries huge number of information which is very useful for users. However it may contain some unnecessary informations for users. The unwanted informations should be cleaned or removed to obtain clean data.

Search Page :

Retreving relevant information from particular web page can be possible by giving search queries. The effective way of navigating web page by search engines and by the customer way can be done with clustered web contant.

Result Page :

The web page visitor Information and the last accurate result obtained from the search are stored in the result page of the content mining.

B. Web Structure Mining

Web information and knowledge about the links between the organizational structure are derived by Web structure Mining. Scientific quotation analysis theory states that the interconnection document data contains a wealth of useful information. Considering the complexity of the structure, ignoring the structure of information the standard search engine takes Web as a collection of documents. The finding of the authoritative pages, center pages are found to improve retrieval performance, the proper guidance from the Mining of the structure and Web page structure is given to the classification and clustering.

Web structure mining contains :

- Links Structure Mining
- Internal Structure Mining

URL Mining Links Structure :

Link Mining is an emerging research area evolved from the researchers interest in the Web Mining, structure analysis researches. Link analysis is a traditional way of link mining research, which is based on link based classification, cluster analysis, link type, link strength and Link Cardinality.

Internal Structure Mining :

The ranking of page and search results information through filtering can be provided by internal Structure Mining. The different Web sites similarity and relationship can be analyzed with this model.

URL Mining :

The Web page can be connected to another different location either in same Web page or to a different web page which is done by hyperlink from the URL Mining.

C. Web Usage Mining

Web usage mining focusses on techniques that could predict user's behavior while the user interacts with the web and also it discovers the meaningful pattern from data generated by client server transaction on one or more web localities [1]. A web is a collection of inter-related files on one or more web servers. Data mining techniques for pattern analysis can be adapted in web usage mining applications. It generates the data from the server access logs, refers' logs, agent logs, client sides cookies, user profiles, metadata, page attributes, page contents and site structures automatically. It is a technique to predict users' behavior when the user interacts with the web.

Web usage mining can be categorized into three phases:

- Preprocessing
- Pattern Discovery
- Pattern Analysis

Preprocessing :

According to the client, server and proxy server, preprocessing is the first approach to retrieve the raw data from web resources and cleaning the data. It is the transformation of the original raw data into required log of web personalization process [8].

Pattern Discovery :

Necessary patterns for the web personalization process from the cleaned dataset are used for discovering the knowledge and to implement the techniques which will be used for machine learning.

Pattern Analysis :

The necessary patterns gathered after applying certain rules, are analyzed. It checks whether the pattern is correct on the web and guides the process of extraction of the information/ knowledge from the web [18].

Web usage mining is used to retrieve the hidden information in the log files from one or more web pages. The aim of the Web usage mining is to retrieve structure and analyze the behavioral patterns and user profiles of the users who interact with the websites [6], [20]. Web usage mining uses the secondary web data such as web server access logs, proxy server logs, browser logs, user profile registration data, user sessions or transactions, cookies, user queries, bookmark data, mouse clicks and scrolls and click stream data [6]. The knowledge patterns discovered are represented as a group of pages, objects or resources that are continuously used by a community of users with the same interests and needs [4], [16].

The Web usage mining application areas are as follows:

- ◆ Personalization
- ◆ System improvements
- ◆ Web customization
- ◆ Business intelligence
- ◆ Usage characterization

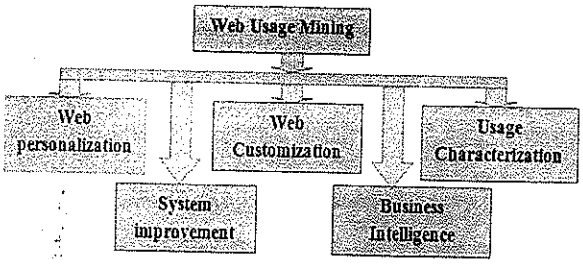

Figure 2 : Web Usage Mining Application

The general architecture for web usage mining shows five different process Fig. 2. Cleaning irrelevant data from web data is the initial stage of every web usage mining, which combine multiple logs and incorporating referrer logs [15]. A single line logical log entries should be partitioned into logical clusters, which has to be done after the cleaning process. The transactions identified are made into clusters for easy categorization of same order transactions [17]. For instance, the format of the data for the association rule discovery task may be different from the format necessary for mining sequential patterns. Finally, a query mechanism will allow the user or an analyst to provide more control over the discovery process by specifying various constraints [16], [5], [2].

III. Web Personalization

Web personalization is the process of customizing or modifying the website according to the needs of each specific user; this can be possible by taking the knowledge from the log files which contain the information about the users' navigational behavior [10]. The usage content data and the usage structure data can be integrated for the better process of personalization [11]. Web personalization creates an important impact on e-commerce. Web personalization proves the easiest way of providing services to e-commerce, e-learning, e-banking, etc [19]. Web mining applications improvement makes the services easy with the users' requirement in an effective and efficient manner [3].

A. Personalization Strategies :

Personalization falls into four basic categories, ordered from the simplest to the most advanced:

Memorization : It is the simplest and most widespread form of personalization. User information such as name and browsing history are stored (e.g. Using cookies) which will be used later to recognize and greet the returning user. It is usually implemented on the Web server [34].

Customization : This form of personalization takes the user's preferences as input from the registration forms in order to customize the content and structure of a web page.

Guidance or Recommender Systems : A guidance based system tries to recommend hyperlinks automatically that are deemed to be relevant to the user's interests, in order to facilitate access to the needed information on a large website.

Task Performance Support : In these client side personalization systems, a personal assistant executes actions on behalf of the user, in order to facilitate access to relevant information.

B. The Web personalization process can be divided into four distinct phases :

A collection of Web data : Implicit data includes past activities/click streams as recorded in Web server logs and/or via cookies or session tracking modules. Explicit data usually comes from the registration forms and rating questionnaires.

5.	www.lumio.com	It is a part of a commercial product Lumio's Re:cognition which provides personalization to the website users based on the current context information, like pages visited, navigation paths and timing information.	An Analytics-based Context Server operator which is a system works with the collected user data, extraction, management and deployment of the user's information. Javascript agents are used for collecting information about identifying users. Context Assembler gathers the information about a visitor's experience from the user information.	Re: action
6.	Ngu and Wu,(1997)	SiteHelper is an another research prototype that offers multiuser guidance functionality by means of static hyperlink recommendation focussing on a certain topic [24].	The SiteHelper tool employs classification techniques to build the set of rules that represent user's interests. After discovering the rules, the system can recommend the web page according to their interest [24].	SiteHelper
7.	Spiliopoulou et al.,(1999)	WUM is a system that offers multi-user customization function by modifying the hyperlinks of a particular Web page to include links to pages that have been visited by customers and not by non-customers. If the visitor is a customer, then the web page remains the same [25].	It uses data from both registered and anonymous users. ORACLE 9i database and uses data combine to form Oracle9iAS. Set of Java API calls are used to capture the navigational behavior of users.	WUM
8.	Schwarzkopf (2001)	The system used for customizing UM2001 conference site, based on the user's interest. A static personalization policy is followed to give guidance and customization at the beginning of a user session and will not give such guidance for the user navigation [26].	The system collects user's information from Web server log files and performs user and session identification by assigning an ID, which is generated by the Web server. This ID is included in every Web page requested by the user and at the same time it is recorded in the log file replacing the IP address. The personalization solution is implemented by building offline the model of each visitor directly from the Web server log files, with the use of simple Bayesian Networks [26].	Schwarzkopf (2001) software for UM2001 for conference site
9.	Mobasher et al. (2000) Yan et al. (1996) and Kamdar and Joshi (2000)	The system used by these three researchers is static function which follows simple multiuser guidance functionality. The navigational behavior followed by the current user is used for the personalization policy [27].	WebPersonalizer (Mobasher et al., 2000): Active session is a short term model used for recording the user navigational behavior for online personalization. The most recent navigation history of the user is given additional weight to the recommendation phase [27]. Cooley et al., 1999: introduced a method for identifying users. Yan et al. (1996) and Kamdar and Joshi (2000): The Clustering techniques used for pattern discovery are similar to Mobasher clustering techniques [28], [29].	WebPersonalizer and Research prototypes that are used to offer simple multi-user guidance functionality.

10.	Pabarskaite and Raudysv(2007)	The system uses users' behavior extraction from Web log and customer information personalization data mining with large-scale Web log mining technique [30].	Clustering technique is used for grouping users, classification of differentiating users, Association algorithm is used in relation between each users' behavior, sequential rules are used for arranging necessary rules for analyzing identified patterns and OLAP applications are used for visualized outcome for analyzed patterns [30].	Analog, freeware of WUM and Commercial software.

Table 2 : Used Measurement for Existing Systems in Web Personalization

S. No	Author	Technique/ Method	Time Measure	Accuracy
1.	Yan et al. (1996)	Minimum number of Pages and minimum cluster size are set to 5 and maximum distance variation is set from 1 to 10 for clustering users from working logs [28].	The time taken for completing the successful clustering ranges from 33 to 60 seconds on DEC Alpha working station [28].	N/A
2.	Kamdar and Joshi (2000)	The distance evaluation used for clustering are users with cookies and without cookies. The separation is done because of unregistered visitor classification. The intra-cluster distance is set to 0 and inter cluster distance is set to 1. The parameters are same for five different sets of log files [29].	The categories done with cookies for obtaining session take less time compared to session obtained without cookies [29].	The Research results show improvement by 20 % for intra cluster distance.
3.	Cyrus Shahabi and Farnoush Banaej (2003)	Pure Euclidean distance (PED) is improvised with Projected pure Euclidean distance (PPED) for clustering the log with FM cluster. The Process is updated with dynamic and real time data collection [31].	The average time taken for viewing the page is improved. The average view time taken is 15 seconds and the accuracy improvement is up to 13% [31].	Projected pure Euclidean distance (PPED) shows a 30 % improvement than Pure Euclidean distance (PED) for clustering. The accuracy of Hit-count is also increased to 40% [31].
4.	M. Baglioni et al. (2003)	The technique is to divide the case of logs into two levels. Level 1 for a single user session and Level 2 is for visiting recent past by the user. The majority classifier algorithm is used before clustering [32].	N/A	The classification accuracy obtained from level 2 is 54.8% (The classification is used for unregistered visitors, so level 2 session is taken for classification) [32].

[24] Ngu,D. S. W. and Wu,X. 1997, SiteHelper: A localized agent that helps incremental exploration of the world wide web. Computer Networks and ISDN Systems: The International Journal of Computer and Telecommunications Networking, 29(8), pp. 1249-1255.

[25] Spiliopoulou, M., Faulstich ,L. C. and Wilkler, K. 1999, A data miner analyzing the navigational behavior of Web users, In : Proceedings of the Workshop on Machine Learning in User Modeling of the ACAI99,Chania,Greece, pp. 54-64.

[26] Schwarzkopf, E. 2001, An adaptive web site for the UM2001 conference, In: Proceedings of the UM2001 Workshop on Machine Learning for User Modeling, pp. 77-86.

[27] Mobasher, B.,Cooley, R. and Srivastava, J. 2000, Automatic personalization based on Web usage mining, Communications of the ACM, 43(8), pp. 142-151.

[28] Yan,T., Jacobsen.,W., Garcia-Molina, M.H., and Dayal, U. 1996, From User Access Patterns to Dynamic Hypertext Linking, WWW5/Computer Networks 28(7-11), pp. 1007-1014.

[29] Kamdar,T. and Joshi,A. 2000,On Creating Adaptive Web Sites using Web Log Mining. Technical ReportTR-CS-00-05.Department of Computer Science and Electrical Engineering University of Maryland, Baltimore County.

[30] Z. Pabarskaite and A. Raudys, 2007. A process of knowledge discovery from web log data: Systematization and critical review, J. Intelligent. Information. System. 28(1). pp. 79–104.

[31] C. Shahabi and F. B. Kashani, 2003. *"A Framework for Efficient and Anonymous Web Usage Mining Based on Client-Side Tracking"*, Published in proceeding of WEBKDD '01 Third International Wprkshop on Mining Log Data Across All customers Touch Points, pp. 113-144.

[32] M. Baglioni , U. Ferrara , A. Romei , S. Ruggieri , F. Turini and Via F. Buonarroti, 2003, *"Preprocessing and Mining Web Log Data for Web Personalization"*, Published in 8th Italian Conf. on Artificial Intelligence, vol(2829), pp.1-12.

[33] Kaikala Anjani Sravanthi and Yalamarthi Madhavi Lata, 2013, *"Web Mining Using Cloud Computing"*, Published in International Journal of Emerging Technology and Advanced Engineering, vol.3(4),pp.139-144.

[34] O. Nasraoui. 2005, World Wide Web personalization. In J. Wang, editor, Encyclopedia of Data Warehousing and Mining, pp. 1235–1241. Idea Group.

[35] Sita Gupta and Vinod Todwal, 2012, Web Data Mining and Applications, International Journal of Engineering and Advanced Technology, Vol.1(3).pp.20-24.

AUTHOR'S BIOGRAPHY

R. Gobinath, completed M.C.A. He is doing Ph.D. (Full Time) in Karpagam University, Coimbatore. He has published two papers in International Journal and presented papers at National and International Conferences. His area of research is Data Mining and Web Mining.

Dr. M. Hemalatha, completed M.Sc., M.C.A., M. Phil., Ph.D (Ph.D, Mother Terasa Women's University, Kodaikanal). She is a Professor and guiding Ph.D Scholars in Department of Computer Science at Karpagam University, Coimbatore. Twelve years of experience in teaching and has published more than hundred papers in International Journals and also presented more than eighty papers in various national and international conferences. Area of research is Data Mining, Software Engineering, Bioinformatics, and Neural Network. She is a Reviewer in several National and International Journals.