A Survey on Web Personalization Approaches

Beenish Siddiqui[1], Dr.Nidhi Tyagi[2]

[1]Student, Department of CSE, M.tech, Meerut Institute of Engg. & Technology, Uttar Pradesh, India.
[2]Associate Professor, Department of CSE, Meerut Institute of Engg & Technology, Uttar Pradesh, India.

---***---

Abstract - *Data on the internet is exponentially increasing and becoming more complex. Retrieving the most relevant information from the internet according to user needs becomes more difficult because of this huge collection of data. To find the relevant information it is mandatory to go through the long list of documents available on the web, but it is time consuming. The approach of Web Personalization meets the user expectations as it personalise the information available on the web according to the user interests and preferences. By considering all the benefits of web personalization, this paper presents an extensive survey on the various approaches proposed by researchers in Web Personalization.*

*Key Words***: Web personalization, Web Usage Mining, User profile.**

1.INTRODUCTION*)*

Web is a large repository of information. The classifying and retrieving the information exactly based on the needs of user is more important. Therefore, the recruitment for predicting the user needs in order to improve the usability and user's retention of a Website can be addressed by personalizing it. Personalization is generally means of meeting the user's need more effectively and efficiently. Users would get things and results according to their interests and expectation without giving much more input.

The process of customizing the content and structure of a Website to the specific and individual needs of each user by taking advantage of user's navigational behaviour is known as Web Personalization [1]. The process of Web personalization starts with the collection of user data, collected data is joined with user needs, demands and result is a specifically designed Website. Some basic elements used to design personalized Web Site are content data, usage data, structure data etc. Website Personalization offers benefits that are a customer-focused and a much comfortable Website, information catered and presented in a much better manner resulting in saving of time, productive results, and higher call-to-action [2].

Web Personalization is viewed as an application of data mining and machine learning to build models of user behaviour that can be applied to the task of predicting user needs and adapting future interactions with the ultimate goal of improved user satisfaction.

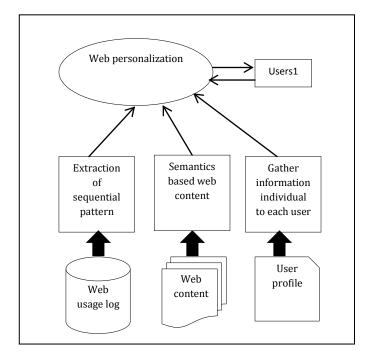## 2. WEB PERSONALIZATION ARCHITECTURE

Figure-1: Architecture for Web Personalization

The architecture for web personalization represented in figure-1; uses Website's content which gives the semantic based content, Web logs created by observing the user's navigational behavior and User Profiles formed according to the user's choices and interests to analyze and mine the information needed for the user to find the pattern expected by the user. This analysis creates recommendations that are presented to the user [2].

3. WEB MINING CLASSIFICATION

Web mining is an important field of data mining used in WWW based information and resources. Mainly data mining brings out the hidden and useful information from the huge web data or resources. Web mining uses data mining techniques to extract useful information from unstructured and huge web data. The unstructured data on the web must be analyzed to increased the Web Personalization System.

Web mining techniques are divided into three categories according to sources of web data. These categories are web content mining, web usage mining, web structure mining [3].

3.1 Web Content Mining

Web content mining describes the extraction of useful information and hidden knowledge from web sources or web content such as text,images and structured record on the web.

3.2 Web Usage Mining

Web usage mining is the application of data mining techniques to bring out interesting patterns from the web usage data such as server logs, client browser logs, proxy server logs, cookies etc.

3.3 Web Structure Mining

Web structure mining is used to identify the arrangement of hyperlink in html document and deduce the knowledge.

4. WEB USAGE MINING

Web usage mining uses web data sources in order to find out hidden knowledge about users. It also recognize the behavior of the user on the web. This process depends on the application of statistical and data mining methods to the Web log data and gives the result in a set of useful patterns that specify users' navigational behaviour[1]. The data mining methods used are: association rule mining, sequential pattern discovery, clustering, and classification. Web usage mining process must follow the three basic steps. These steps are data prepocessing, pattern discovery and pattern analysis of a website.

4.1 Data Preprocessing

Data preprocessing is the vital part of web usage mining. Before the process begins the prepration of data is done by; firstly logging of the data; performing accuracy check; integrating data available at different sources; transforming the data into a session file; and finally structuring the data as per the input requirments [3]. Data preprocessing task is done in following steps:

Data cleaning is the process where irrelevant data is removed from the database. Irrelevant data comes from non-analyzed source, data with missing attributes or the attribute that are not needed for the project goal. This process helps in reducing the size of data and any false association that can be generated because of redundent.

User identification is the next step in data preprocessing. This is used to identify all the unique users from the data log. Certain heuristics can be used to identify the users uniquely. Suppose, if the dat log contains common IP address to the user, still there is need to differentiate the agents by obtaining different user sessions from which the request have been made. With the help of the referrer log from the data, different access paths can be generated by taking the respective users. Simple heuristics may not always provide correct information. For example, same browser agent is used by two users having the common IP address and then both the users appear to be as a single user. There can also be the exactly opposite scenario, where the same user can have a different IP address and be using a different browser resulting in confusion again [3].

Session identification defines the number of times the user has accessed a web page [4]. Time out mechanism can be used to identify the accessed time of the user for a respective web page. Time limit for the access of a particular page is defined by this mechanism and usually this limit is of 30 minutes, if this limit exceeds 30 minutes then this will be divided into more than one sessions. This mechanism faces the problem of chaching from browsers and produces incomplete web logs [3].

Path Completion identifies from where the request came, what all pages are involved in the path from the start till the end. The referrer helps in determing the path for a particular request. This process solves the problem of missing entries that mislead in tracking the request [4].

Data summarization, after all the above process data summarization is performed. The data is inserted into a relational database system for further generalization and computations which helps in reducing dimentionality of the data.

After applying all the above processes, final data is produced should be flawless and ready for data mining to contribute in producing correct and effective results.

4.2 Pattern Discovery

User access patterns are discovered from the user access logs, referrer logs, user registration logs etc. Discovery of user access patterns is the main goal of Web Usage Mining activity. It applies statistical and data mining techniques to

the preprocessed Web log data to disscover useful patterns. There are numbers of methods and algorithms. Same are discussed below:

Association rule mining is a technique for finding frequent patterns, associations, and correlations among sets of items and are used in order to expose correlations between pages accessed together during a server session [1]. **Sequential pattern discovery** is an extension of association rules mining in that it tells patterns of cooccurrence incorporating the notion of time sequence. Patterns might be a web page or a set of pages accessed immediately after another set of pages. Using this approach, useful users' trends can be discovered, and predictions concerning visit patterns can be made [1]. **Clustering** is the task of grouping a set of items in such a way that items in the same group are more similar to each other than those in other groups[1]. **Classification** maps the data item into one of several predetermined classes [1].

4.3 Pattern Analysis

Once the patterns are discovered from usage data, a further analysis has to be conducted. Depending on the techniques previously used, the exact methodology shuold be applied for analysis. Mostly used methods of analyzing such patterns are either by using a query mechanism on a database where the results are stored, or by loading the results into a data cube and then performing OLAP operations. Additionally, visualization techniques are used for an easier explanation of the results. Using these results in association with content and structure information concerning the Web site there can be extracted useful knowledge for modifying the site according to the correlation between user and content groups [1].

5. USER PROFILING

User profiling helps the system to distinguish between different users or groups of users. The objective of user profiling is the creation of an information base that contains the preferences, characteristics, and activities of the users [1].

It is a challenging issue for web users to collect useful information from the web. Now days, the web users want more smart systems to collect useful information from the huge collection of web data to meet their information needs, therefore user profiles are created for user background knowledge description [5][6][7]. The personal data related to a specific user is assembled and description of the

characteristics of a person can be represented on the user profile. It can also be described as the computer representation of a user model [8]. User profiles are divided into three groups: Interviewing, semi-interviewing, and non-interviewing.

Interviewing user profiles, also called manual user profiles and are developed by using manual techniques, such as questionnaires, interviewing users, etc. These profiles are considered to be perfect user profiles.

Semi-interviewing user profiles are developed by semi-automated techniques with limited user involvement. In these techniques, a list of categories is provided to the users and asks for interesting and non-interesting categories.

Non-interviewing techniques, discovers user interests and observe user activity and behaviour on the web. In these techniques user involvement is not required.

5.1 Techniques Using User Profile

Content-based filtering: This technique is applied to a textual document which results document's relevance by matching the keyword contained in user profile with the keyword extracted from the text document [9]. This is the most common way to use the user profile that enables personalization on an individual basis. User profile are stored at server therefore they cannot be transmitted through network [8][10].
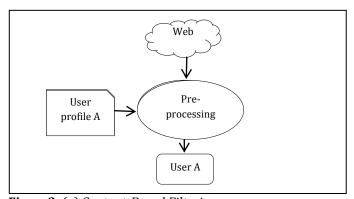

Figure2: (a) Content-Based Filtering

Collaborative filtering: In this technique user profiles of a group of people are collected and based on similarities of the collected profiles generates recommendations as given in the figure 2(b). This technique is based on the assumption that users with similar behaviour are grouped together. User profile must be compared and placed at server [8][10].
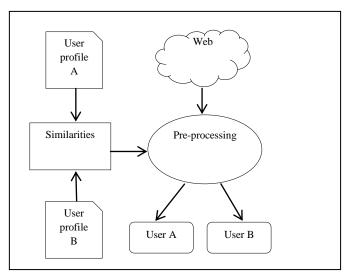
Figure-2: (b) Collaborative Filtering

User profile sharing: In this technique user profile is shared between different personalized apllication that want the same content of user profile. This association enables to gain a much more knowledge about the user's interest. User profile are stored at the browser because all the personalized web applications need to have access to the complete set of profiles for a individual user [8][10].
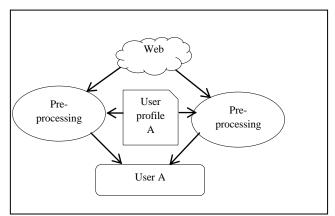

Figure-2: (c) User Profile Sharing

6. CONCLUSION

World Wide Web is the largest source of information. To reduce the information overload Web Personalization is an important tool that displays the information exactly needed by each user. This paper describes the web usage mining and its utilization in web personalization and the basic steps of WUM. Various approaches carried out to improve the performance of Web personalization is optimized approach for preparing user profile should be worked upon. User data submitting in different format should be searched and analysed.

REFERENCES

[1]   Magdalini Eirinaki and Michalis Vazirgiannis, "Web Mining for web personalization", ACM Transactions on Internet Technology, 03(01):1-27, February 2003.

[2]   Beatric ,Ryan Fernades, Leo.J.Peo, Nikhila Kamat, Sergius Miranda, "New Approaches To Web Personalization Using Web Mining Techniques", International Journal Of Computer Science And Information Technologies,Vol.5(2), 2195-2201,2014.

[3]   S.Jagan, Dr.S.P.Rajagopalan , "A survey on Web Personalization of Web Usage Mining", International Research Journal Of Engineering And Technology(IRJET), Volume:02 Issue:01|March-2015.

[4]   Sasa Basnjak, Mirjana Marie ,Zita Basnjak, " The Role Of Web Usage Mining In Web Application", Management Information Systems, Vol 05(01): 31-36,2010.

[5]   S. Gauch, J. Chaffee, and A. Pretschner, "Ontology-Based Personalized Search and Browsing" Web Intelligence and Agent Systems, vol. 1, nos. 3/4, pp. 219-234, 2003.

[6]   Y. Li and N. Zhong, "Web Mining Model and Its Applications for Information Gathering" Knowledge-Based Systems, vol. 17, pp. 207-217, 2004.

[7]   Y. Li and N. Zhong, "Mining Ontology for Automatically Acquiring Web User Information Needs" IEEE Trans. Knowledge and Data Eng., vol. 18, no. 4, pp. 554-568, Apr. 2006.

[8]   K. Sridevi1 And Dr. R. Umarani, "Web Personalization Approaches: A Survey", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 3, March 2013

[9]   Morita M., Shinoda, Y., "Information Filtering Based on User Behaviour Analysis and Best Match Retrieval", in Proceedings of the 17th International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp. 272-281, 1994.

[10]   Doddegowda BJ, Dr. GT Raju, "A Novel Approach for Web Personalization through Web Mining Techniques", An International Journal Of Advanced Computer Technology 4(5), May-2015.