Web Personalization Systems and Web Usage Mining: A Review

Rajesh Shukla
PhD Scholor
RGPV Bhopal

Sanjay Silakari,Ph.D
Professor and Head
Deptt of CSE, UIT RGPV, Bhopal

P K Chande,Ph.D
Group Director
Truba Group of Institutions, Indore

ABSTRACT
Nowadays, the field of web personalization is growing exponentially. From e-mail, e-trading, internet forum to social networking based websites, directly or indirectly utilize web personalization and recommendation system for providing customized services to their loyal users. Personalization is achieved through web mining, i.e. extracting knowledge from the collected data. Knowledge is then filtered and processed to model user behavior that forms the basis of a personalized system. This paper presents a brief review of recent research efforts in web personalization and recommendation by means of web usage mining, for the benefit of research in this area. It also elaborates the role of web usage mining in personalization, and presents the open challenges that are yet to be met.

General Terms
Web Usage Mining

Keywords
Web Usage mining, Web Personalization, Recommendation System

1. INTRODUCTION
Personalization is a relatively new and challenging field for web content delivery. In order to meet expectations of visitors, customers and loyal users, web world is struggling to offer excellent customized services during their interaction with the system. The impact of personalization and recommendation system can be experienced by the rapid popularity that this area has gained in the last few years. Customers preferably choose to visit those websites, which understand their needs, provide them rapid value added customized services and easy access to required information in simple understandable format. Web personalization and recommendation system plays a major role in meeting this goal.

Corporate world looks towards the huge volume of transactional and interaction data generated by the Internet for R&D that facilitates the creation of new innovative competitive services and products [3]. In today's e-business world, most of the major e-commerce players have adapted the web personalization and recommendation system including Yahoo!, Amazon, eBay, Netflix, NewsWeeder, IBM and many more.

A recommendation system learns from a customer's behavior and recommends a product in which users may be interested. It helps to build up a long lasting relationship with loyal users of website. Various vendors offer web personalization tools that can be employed in existing systems to achieve personalized web system. IBM Product Recommendations tool helps to present the most relevant, effective and timely recommendations wherever customers are in the buying process [17]. Adobe (Omniture) Test & Target is an optimization tool that Identifies and executes unlimited A/B and multivariate tests (MVTs) whenever required. It also measures the effectiveness and relevance of contents across any online channel and increase content relevance through segmentation, targeting, and automated personalization [13]. BT Buckets is a free personalization tool and behavioral targeting tool that can be integrated with Google Analytics. Another tool Magiq Dynamic Personalization allows marketing teams to conduct website content personalization campaigns, and optimize their digital marketing communications to the individual customer [15]. The WP Greet Box plug-in enables a website to display different greeting message to each visitor depending upon the web pages visited by him previously. ATG's eStara enables customers to initiate phone calls with the help of agents that assist customers with their order.

In active personalization, users explicitly supplies information to system in order to get customized services/features. Once experiencing the benefits, users may be more willing to surrender information without caring of its consequences. Moreover, in passive personalization, user is often unaware of what information is being captured. Since personalization is achieved by means of intensive information about users, hence privacy standards like P3P [16] must be employed. However we will not discuss on security issues in web personalization any more, as it is beyond the scope of this paper.

This article is organized as follow: section.2 provides an overview of web data and its types, introduction to web mining and its types, and limitations of various methods for analyzing web data. Section.3 explains the steps involved in web usage mining i.e. data collection, data processing, knowledge discovery and knowledge post-processing. In this section, various approaches for acquisition of data and their drawbacks are also discussed. Section.4 outlines the recent noteworthy contributions in the field of web personalization and recommendation through web usage mining. Section.5 presents the challenges and problems that personalization systems face. Finally, section.6 concludes the paper.gutter.

2. HOW WEB PERSONALIZATION IS ACHIEVED?
Web personalization and recommendation system utilizes data, which is collected explicitly or implicitly during the interaction of user with website. Such a collection of data is known as Web data that can be divided into four major categories: Content data, Structure data, Usage data and User profile data as shown in Fig.1.

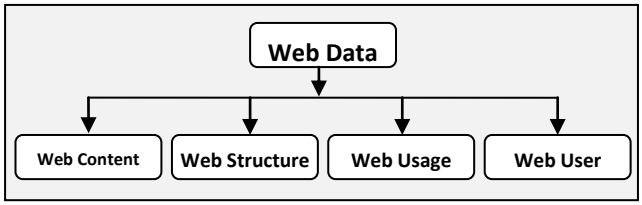

Fig.1. Classification of Web Data

Content data is intended for end-user in simple text format, images or structured information retrieved from databases. Structure data represent how the contents are organized internally. It may include data entities used in web pages, such as HTML or XML tags, and hyperlinks to interconnect web pages. Usage data represents usage of website, it consists of visitors' IP Address, time and date of access, complete paths of files or directories accessed, and other attributes that can be included in a Web access log [26]. User profile data comprises of personal information of each website user, such as name, age, sex, country, qualification etc. and information about users' interests and preferences. It is explicitly provided by user during the interaction with website while filling registration forms or questionnaires.

Once the web data is procured, task of personalization can be initiated. The process of web personalization includes identifying the visitors of website, retrieving their profile data, selecting contents that suit to their profile, and then displaying those contents in most pleasing and easily understandable format. The method of discovering useful knowledge from collected web data is known as Web Mining [30], which can take any of the following three forms as shown in Fig.2: Web Content Mining, Web Structure Mining, and Web Usage Mining [4, 32].

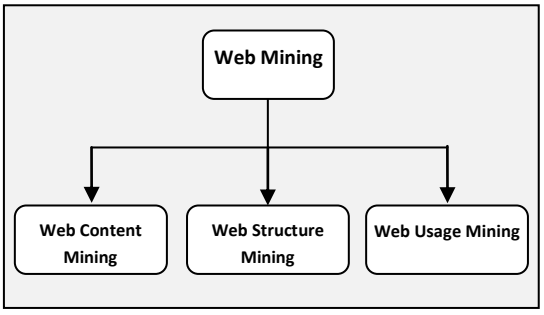

Fig.2. Classification of Web Mining Techniques

Web Content Mining finds application of data mining techniques for extracting knowledge from content data. Web Structure Mining is the application of data mining techniques to structure data, so as to discover the model of link structure in web pages. Lastly, Web Usage Mining utilizes aims at discovering interesting patterns in usage data by using data mining techniques. In this paper, we are more interested in personalization system through web usage mining, which plays a major role in predicting user interests and behavior. These predictions act as foundation of personalization and recommendation systems.

To analyze the web data, various approaches have been proposed by researchers across the globe in past few years: Magdalini Eirinaki et al. [26] classified these approaches into four major classes: Content-based filtering systems, Social or Collaborative filtering systems, Manual rule-based filtering systems, and Web usage mining based systems. Content-based filtering systems model the behavior of individual user based on his past interests, personal preferences, and browsing behavior. Once user modeling is completed, system starts recommending items (to users) that match individual user's profile.

The goal of Collaborative filtering systems is to achieve personalization without analyzing the web contents. Such systems invite users to explicitly rate the available items, or reveal their personal preferences which are recorded by the system. The system then performs categorization of users, based upon the information provided by them. Such recommendation engines work on assumption that the users with similar behavior have correlated interests. Thus depending upon the category in which a user falls, system can acquire the information that a particular user may be interested in.

Manual rule-based filtering system requires manual intervention of website designer and user's co-operation in order to achieve personalization. Under this approach, a set of questionnaires derived from a decision tree is presented to users. Based on the answers given by user, a set of rules are defined manually, and a static user model is created. Depending upon the underlying rules and user model, contents of web pages are tailored according to user's needs. Yahoo! personalization engine and IBM Websphere are the examples of rule-based filtering system. A combination of content-based, rule-based and collaborative filtering techniques can also be employed for achieving more accurate results [10]. The approaches discussed above have their own limitations, which are briefly summarized in Table.1.

Table.1. Limitations of various approaches for analyzing Web data

Approach for analyzing Web data	Limitations
Content based filtering	• very complex to implement • employs NLP and IR, which are still under development
Collaborative filtering systems	• do not scale well to large numbers of users • generates poor quality of recommendations for those users who have rated small number of items
Rule based filtering	• requires web designer and user intervention • requires significant efforts in construction and maintenance

In this paper, we concentrate on web usage mining based systems that focuses on discovering interesting patterns from usage data. It is now widely recognized that usage mining is a valuable source of ideas and solutions for web personalization [10]. The web usage data represents details of user-website interaction, which is appropriate to create user model that represents user's behavior, interests and personal preferences. The constructed user model can be used by personalization system in fully automated way without any human interference, for carrying out the personalization task.

3. WEB USAGE MINING IN WEB PERSONALIZATION

When data mining techniques are applied on web usage data in order to extract useful knowledge regarding user behavior, it is known as web usage mining. It is an approach for collecting and preprocessing web usage data, and then constructing models that represent the behavior and interests

of users. Such models can automatically be used by personalization system for predicting user's personal interests and thus enhance his surfing experience with the website. The web usage mining involves collection of data from various sources, pre-processing of collected data, discovering useful knowledge, and finally post processing the knowledge.

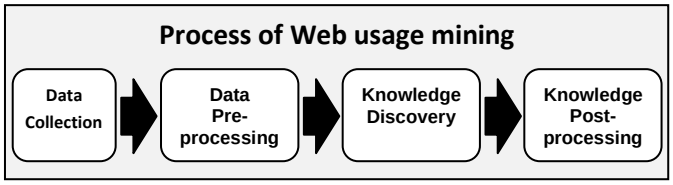

Fig.3. Overview of web usage mining process

i. Data Collection

In this stage, the usage data is collected from a range of possible sources after which, their contents and structure is recognized. Data is collected from a various sources such as those from web servers, clients connected to server, intermediary sources, and the third-party databases [43]. A web server can contribute to data collection through server log files, dispensed cookies on client side, explicit user input, and external data.

Server logs are considered as a major data source for discovering knowledge about website usage, but it cannot be fully trusted because servers store cached copy of accessed web pages for a certain time period. Firstly, when a request arrives for same web page then these cached copies are presented before users without creating its entry in the server log. Although cache control response headers resolve this problem but it diminishes the sole purpose behind caching. Secondly, use of proxy server hides the IP-Address of hosts and assigns them same IP-Address. Thus web server creates log entries with same IP-Address even when requests were coming from multiple hosts. This leads to the problem called IP misinterpretation.

A cookie is Unique ID generated by web server that is copied as a small file (<= 4KB) on client machine, server also records it for identifying the client later. Cookies are also capable of storing other usage data, but their small size delimits the possible benefits. When client accesses the same website again from same machine, browser reads that Unique ID and sends it to server. Thus server identifies its user with the help of Unique ID that was assigned to user when he visited the website last time. Cookies suffer from two major problems; firstly this approach fails if client disables the cookies in its browser. Secondly since cookies are intended to work with browser on a host, therefore server may misinterpret a user when several users use same computer and visit the same website.

Explicit User Input data is explicitly provided by users by means of registration forms. But this approach cannot be considered as good because firstly, it incurs extra burden on user, and it discourages users to visit the website. Secondly, information produced through this approach cannot be trusted because users tend to reveal minimum possible personal information due to privacy issues. External data is the data obtained from third party, who maintains user information in its database. But this approach may not suit to privacy and security norms in several countries.

Client Side data Collection is considered more reliable than using server side sources as it overcomes both the caching and session identification problems [43]. Along with log files intentional browsing data from client side like “add to my favorites”, “copy” is also added for efficient web usage mining [47]. Client side data are collected from the hosts who visit the Web site. One of the most common techniques for acquiring client side data is to dispatch a remote agent, implemented in Java or JavaScript [8]. These agents are embedded within the web pages generally by using Java applets or JavaScript. However, this approach has its own problems, such as data collecting agents may degrade the system performance on client side. Furthermore, these methods require co-operation of users, who may not allow an agent running on their side. Web users often activate security mechanisms that restrict the use of such agents without user's permissions. Since this approach may easily be misused to compromise user's security and privacy, hence it is difficult for users to accept it.

A proxy server is an intermediate server between client and web server, which runs proxy software. It ensures security and privacy, and facilitates administration over internet activities within an enterprise. Like web server, a proxy server also caches web pages for controlling network traffic [28]. It also maintains access logs, with similar format to that of web servers in order to record the requests to and responses from web servers. It may be seen as valuable data source for data related to a group of users who access web server through a common proxy server. But like web servers, problem of web page caching and IP address misinterpretation also applies to data collected from a proxy server.

Packet sniffer can either be a software or hardware device used to monitor network activities. It is capable of collecting and analyzing the usage data in real time. It can reveal the information not contained in log files, such as complete Web page that has been requested can be included in the sniffed data [2], whether user cancelled a request in midst of an operation by pressing browser's stop button etc. But information obtained through this approach has few major disadvantages as compared to log files. Since data is not logged but collected in real time, hence it may be lost forever and additionally, packets may not arrive in the same order in which they were sent. Also TCP/IP packets are increasingly transmitted in encrypted format, which further reduces the ability of packet sniffers to extract useful information. Finally they are considered as a severe threat that can compromise user's privacy and security policies. Fig.4 presents various sources for collecting usage data, and major drawbacks of these sources are summarized in Table.2.

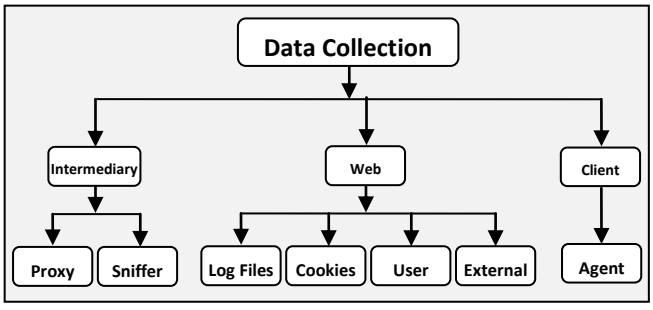

Fig.4. Sources for collecting web usage data through web server

Table.2. Drawbacks of various data sources

Source	Drawbacks
Web server	
Server log files	• Web-cache • IP-Address misinterpretation
Cookies	• User misinterpretation • May be disabled due to security and privacy issues
Explicit user input	• Discouraging from user's point of view • User may not provide correct and sufficient data
External data	• Faces legal obstacles • Privacy may be compromised
Client side	
Software agent	• May degrade system performance on client side • Very hard for users to accept such agents on their end
Intermediary	
Proxy server	• Web-cache • IP-Address misinterpretation
Packet sniffer	• Since data is not logged and hence it may be lost forever • TCP/IP packets may not arrive in sequence • Information cannot be obtained from encrypted packets

ii. Data Pre-Processing

The data collected during first stage is usually diverse and voluminous. Therefore it is necessary to preprocess it by filtering unnecessary and irrelevant data, predicting and filling in missing values, removing noise, transforming it into more useful format, and resolving the inconsistencies. In Web usage mining, this stage includes identifying the users and their sessions. In addition, it is also necessary to provide a good trade-off between insufficient preprocessing and excessive preprocessing [25]. Data pre-processing consists of Data filtering, User identification, and User session identification as shown in Fig.5.

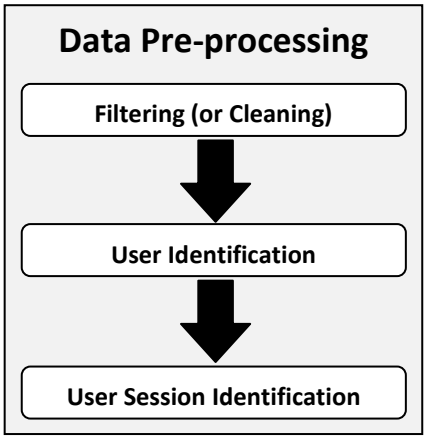

Fig.5. Steps involved in data pre-processing

Data filtering (or data cleaning) is the first step in data preprocessing, where the major task is to clean and filter the raw web data. During this step, the available data is examined, and irrelevant or redundant items are removed from it. Data generated by client-side agents does not require cleaning as it is intentionally collected by the software agents. Since data preprocessing task is domain-dependent hence depending upon the domain of website, it is necessary to differentiate between relevant and irrelevant data, otherwise it may cause loss of valuable information. The records created by spiders and crawlers are also deleted during filtering since it is not considered as usage data.

User identification is the most important step in data pre-processing as without identifying the users who access the website, personalization cannot be achieved. Currently a user informs the system explicitly about who he/she is, by entering a pair of username and password. But a lot of research is being conducted for automating the process of user identification. The simplest strategy is to assign new User ID to each request coming from a different IP address. But this approach cannot be employed in presence of proxy servers between user and web server. However this problem can be rectified with the help of software agents.

User session is a delimited set of web pages visited by a particular user in single visit to the website. Identification of user sessions has also received significant attention as it reveal the navigational behavior and surfing habits of a user, which forms the foundation of personalization system. A user may have a single or multiple sessions during a particular time period. Once a user is identified, the click stream of each user is segmented into logical clusters and this process is known as sessionization or session reconstruction. Various heuristic methods have been used for identifying user sessions.

Spiliopoulou [29] divides these methods into time-based and context-based. In time-based approach, a page viewing time is defined, and a single user session consists of all those web pages, which are requested by a particular user within the page viewing time. However, this heuristic is not very reliable as the exact actions performed by a user may vary greatly and are not known.

A transaction is defined as a subset of user session containing homogenous pages. Cooley et al. [31] made the assumption that transactions have a close relation to browsing behavior of users, and therefore can be identified using contextual information. Based on this assumption, web pages of a website can be divided into navigational pages (that contain primarily hyperlinks to other web pages and are used just for browsing purpose), content pages (that contain the actual information of user's interest) and hybrid pages (which are combination of both types). Context-based classification is not very strict, and depends on the users' perspective. A web page, which is a navigational page for one user might be a content page for the other.

iii. Pattern Discovery

This stage employs machine learning and statistical methods on pre-processed web data in order to extract the patterns of website usage. There are four major machine learning approaches that are often found in literature: clustering, classification, association discovery and sequential pattern discovery [10]. Unlike data pre-processing, the methods employed for pattern discovery are domain-independent, which means same pattern discovery can be method can be applied to websites of different domains without any modification. In this section, we present the application part of various pattern discovery methods without getting into the

details of their algorithmic part. Fig.6 depicts major approaches used to carry out the task of pattern discovery.

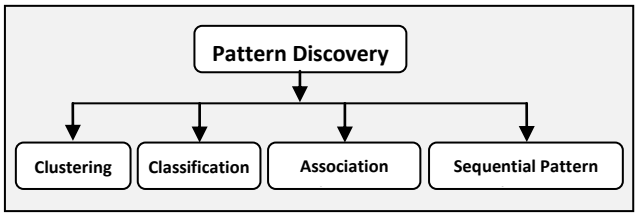

Fig.6. Major Techniques used in pattern discovery

Clustering is a technique to group those users who exhibit similar browsing patterns, or web pages which exhibit similar contents [20]. This approach is employed in majority of pattern discovering methods. Web usage mining allows the overlapping of clusters [11]. It is very important in web personalization because a user or a web page may not necessarily belong to a single group. According to [19] clustering methods can be classified as: Partitioning methods, Hierarchical methods, and Model-based methods. First approach is used to break up a given data set into n clusters (groups). Second approach is used to decompose a given data set into a hierarchical structure of clusters, and third approach is used to find the best match between a given data set and a mathematical model. So far, a number of clustering approaches have been proposed by various researchers, which are briefly discussed in section.4.

Classification of pre-processed data involves assigning a web page or a user to one or more predefined classes. It also helps to develop a profile for items belonging to a particular group according to their common attributes. It is a supervised learning problem [40] where a set of labeled data acts as target vector for training a classifier that can be employed for labeling future data. Based on the organization of classes, classification can be divided into flat classification and hierarchical classification [45]. In flat classification, categories are considered parallel whereas in hierarchical classification, categories are organized in a hierarchical tree-like structure. Majority of existing web classification approaches focus on flat classification. Research specifically on hierarchical web classification is comparatively scarce. The major classification approaches found in literature includes Decision tree induction, neural networks and Bayesian classifiers.

In the context of Web mining, association discovery determines the correlations among each client's accesses to various resources available on server. A transaction is comprised of a set of URLs accessed by a client in one session. Let A and B denote two items, then the union of A and B (AUB) is called an itemset. Association rules are used to calculate the possibility of occurrence of B, given A. The selection of an association rule is based upon two parameters:

- Support: frequency of the itemset in dataset
- Confidence: Conditional predictability of B, given A. It is calculated according to following formula:

$$Confidence = \frac{Frequency\ of\ (A \cup B)}{Frequency\ of\ A}$$

One of the most popular algorithms to find the association rules is Apriori algorithm [33].

The goal of sequential pattern discovery is to identify those event sequences, which frequently occur in the dataset. It is useful in identifying the navigational patterns of users. Under this approach, two types of method are most popular: deterministic methods and stochastic methods. Deterministic methods employ recording of navigational behavior users, whereas stochastic methods analyze the sequence of web pages visited by users for predicting the web pages that the users may be interested to visit in future.

iv. Knowledge Post-processing

To achieve personalization, it is essential to apply post-processing to the knowledge obtained after the passage of raw data through previous four stages. The outcome of post-processing is used by human experts who act accordingly to accomplish the personalization task. So far, there are three major approaches for applying the post-processing to knowledge: Report generation, Query mechanism, and Visualization. Among these, Query mechanism seems to be better than other two approaches as it facilitates the access to frequent itemsets by means of SQL queries as well as visualization mode.

Next section summarizes recent noteworthy contributions found in literature, in the area of web personalization and recommendation system through web usage mining.

4. EXISTING WORK

With the dramatic increase in the number of websites on the internet, tagging has become popular for finding related, personal and important documents. In [Onur Yılmaz, 2013], a tag-based website recommendation method is presented, where similarity measures are combined with semantic relationships of tags. This approach performs well in recommending new websites or catching user's current interests. However, there is no control on the tags provided by users in this system. Although users do not intend to mislead the method while tagging websites, different purposes of tagging can create such a problem.

[A. Vaishnavi, 2012] proposed a technique for developing web personalization system using Modified Fuzzy Possibilistic C Means (MFPCM). The author claims that this approach raises the possibility that URLs presented before a user will be of his interest. [Dimitrios Pierrakos et al., 2012] presented a system that builds and maintains community web directories by employing a web usage mining framework that offers a range of personalization functionalities. It was named as OurDMOZ, which includes adaptive interfaces and web page recommendations. [Bin Xu et al., 2012] extends the traditional clustering collaborative filtering models. They formulate the Multiclass Co-Clustering (MCoC) problem and propose an effective solution to it in order to find meaningful subgroups. They also propose a unified framework that extends the traditional collaborative filtering algorithms by utilizing the subgroups information, for improving their top-N recommendation performance.

So far, collaborative filtering has been the most successful technique in the design of recommender systems, where a user is recommended those items, which people with similar tastes and preferences liked in the past. Although the studies of tag-aware recommender systems have achieved fruitful goals, but there are still few challenges that are yet to meet, which are highlighted in [Zi-Ke Zhang et al., 2012]. [Tamas Jambor et al., 2012] applied the principles of modern control theory to recommendation system, and suggested how to construct and maintain a stable and robust recommender system for dynamically evolving environments.

[Joseph A. Konstan et al., 2012] have shown that the embedding of the algorithm in the user experience dramatically affects the value of the recommender to user.

They argue that evaluating the user experience of a recommender requires a broader set of measures than have been commonly used, and suggest additional measures that have proven effective. [Vivek Arvind. B et al., 2012] proposed an intelligent recommendation system that utilizes the boosted item based collaborative filtering (for the efficient rating of predicted items) and association rule mining technique (for making a personalized recommender system for the target user). This system improves the overall web recommendation precision. [K. Vinodh et al., 2012] used consumer acceptance and use of Information Technology as a model for studying consumer acceptance of an online government service channel. They followed randomized experimental design for testing how web personalization moderates the impact of performance expectancy, effort expectancy, facilitating conditions, hedonic motivation, price and habit on consumer's intention to use a technology.

[Ronaldo Lima Rocha Campos et al., 2011] proposed a multi-agent based system application model for indexing, retrieving and recommendation learning objects that are stored in different and heterogeneous repositories. In order to improve the accuracy, they have come up with an information retrieval model, which is based upon the multi-agent system approach and an ontological model.

Collaborative Filtering is the most popular recommendation technique. However, Classical Collaborative Filtering systems use only direct links and common features to model relationships between users. [Ilham Esslimani et al., 2011] presented a new densified behavioral network based collaborative filtering model (D-BNCF), based on the BNCF approach, which uses navigational patterns to model relationships between users. This approach achieves a high precision when new links are exploited to compute the predictions.

Recommendation services play a vital role in E-commerce. After analyzing the feasibility to combine Case-based Reasoning (CBR) and web log mining with recommendation system in E-commerce, [Ya-min Wang et al., 2011] integrated the two techniques into E-commerce recommendation system effectively. CBR is a paradigm that takes advantage of knowledge obtained from past experience and the state of a specific problem. It uses past similar cases and their solutions in the state of the new problems for solving them. In this work, they proposed a framework for recommendation system that is closer to human thinking mode.

[Samira Khonsha et al., 2011] suggested a framework for web mining-based personalization that combines web usage data with web content, and site structure for predicting user's future requests more accurately. [Utpala Niranjan et al., 2011] proposed an algorithm called modified IncSpan for effective mining of sequential patterns from the incremental database. This algorithm is capable to discover sequential patterns from incremental database based on the sequential patterns obtained from the insert and append database, and the closed sequential patterns are obtained from the resultant sequential patterns.

A challenging problem in recommendation systems deals with unvisited or newly added pages. [Rana Forsati et al., 2009] addresses this problem by introducing a novel Weighted Association Rule mining algorithm. This method can improve the overall quality of web recommendations. [Michael Chau et al., 2007] proposed a machine-learning-based approach that combines web content analysis and web structure analysis. Each Web page is represented by a set of content-based and link-based features, which can be used as the input for various machine learning algorithms. This approach outperforms the traditional text classification methods.

[Karen H. L. TsoSutter et al., 2008] proposed a standard technique that allows tags to be incorporated to standard collaborative filtering algorithms. Firstly it reduces the three-dimensional correlations into three two-dimensional correlations, and then applies a fusion method to re-associate them. The fusion method outperforms standard baseline models with the incorporation of tags. Fuzzy clustering techniques are found to be very efficient in clustering accuracy.

Collaboration filtering techniques used for recommendation require accumulation of vast amount of historical user-preference information, which is queried to provide a personalized experience. Model-based collaborative filtering techniques are preferred over the somewhat more accurate memory-based collaborative filtering techniques primarily due to their higher efficiency and scalability. [Bhushan Shankar Suryavanshi et al., 2005] introduced a web usage profile maintenance scheme called incremental Relational Fuzzy Subtractive Clustering (RFSC), which can efficiently add new usage data to an existing model overcoming the expense associated with frequent remodeling. They also introduced a quantitative measure, called impact factor. When the value of impact factor exceeds a predefined threshold, remodeling is recommended.

5. CHALLENGES IN WEB PERSONALIZATION

Recent advances in web personalization and recommendation system has brought the researchers together from all over the world, for resolving the challenges that are yet to be met. This section presents the major issues that demand rigorous research in this area.

The most important issue to be considered during user profiling is privacy violation. Many users are reluctant to provide their personal information either implicitly or explicitly (such as those obtained from registration forms). Aware users also hesitate to visit those web sites, which make use of cookies or agents. In such an environment, it becomes difficult to achieve personalization.

In data pre-processing phase, the web log data may need to be cleaned from entries of pages that returned an error or graphics file accesses. For some domains, such information may be important whereas for other domains, same data should be eliminated from a log file. Thus overcoming the domain dependence problem is important so that a standard technique can be employed for all the domains.

Another problem to be met has to do with web caching. Accesses to cached web pages are not recorded in the server log and hence such information is missed. Identification of a user from server log entries is also very important but this task becomes difficult in presence of a proxy server between web server and the host.

A visitor having no previous interaction with the website poses a problem to the personalization system as there is no data available for personalizing the interactions with such a user. A similar problem occurs for a newly added item. Due to the absence of rating history, system cannot recommend a new item to users until its rating history has been collected.

Many personalization systems use static profile of users. It should be remembered that user's interests are not static (fixed) and it can keep changing with respect to time, which supports research in dynamic profiling of users. Those personalization systems, which highly depend upon item ratings provided by users, are vulnerable to receive false ratings from users who have their vested interest in making an item popular among visitors.

Although research is already being conducted to resolve these issues since past few decades and as a result, researchers have come up with a range of solutions. But the growing size of data and rapidly increasing demand for personalization in various contexts is continuously posing new challenges to investigators.

6. CONCLUSION

This paper presented the basic concepts involved in web personalization by means of web usage mining. Challenges and problems that need focus of researchers and a brief review of noteworthy recent research efforts are also addressed in this paper. This article can contribute to make people aware of web usage mining, the underlying process in web personalization, and burning issues for further research in this field.

6. REFERENCES

[1] A. Vaishnavi, "Effective Web personalization system using Modified Fuzzy Possibilistic C Means," Bonfring International Journal of Software Engineering and Soft Computing, Vol.1, Special Issue, 2011, pp.1-7.

[2] Anja Feldmann, "Continuous online extraction of HTTP traces from packet traces," In Proc. of World Wide Web Consortium Workshop on Web Characterization, 1998.

[3] Anup Prakash Warade , Vignesh Murali Natarajan and Siddharth Sharad Chandak, "How to Develop Online Recommendation Systems that Deliver Superior Business Performance," Cognizant 20-20 Insights.

[4] Bamshad Mobasher, Robert Cooley and Jaideep Srivastava, "Automatic personalization based on Web usage mining," Communications of the ACM, Vol.43, No.8, 1997, pp.142-151.

[5] Bhushan Shankar Suryavanshi, Nematollaah Shiri and Sudhir P. Mudur, "Incremental Relational Fuzzy Subtractive Clustering for Dynamic Web Usage Profiling," In Proc. of ACM Workshop on Knowledge Discovery in the Web, 2005.

[6] Bin Xu, Jiajun Bu, Chun Chen and Deng Cai, "An Exploration of Improving Collaborative Recommender Systems via User-Item Subgroups," In Proc. of the IEEE 21st international conference on World Wide Web, ACM, 2012, pp.21-30.

[7] Byeong Man Kim, Qing Li, Jong-Wan Kim, and Jinsoo Kim, "A New Collaborative Recommender System Addressing Three Problems," Springer, LNCS, Vol.3157, 2004, pp 495-504.

[8] C. Shahabi, A. Zarkesh, M. Abidi, J. and V. Shah, "Knowledge discovery from users Web-page navigation," In Proc. of IEEE 7th International Workshop on Research Issues in Data Engineering, 1997, pp.20-29.

[9] Dimitrios Pierrakos, Georgios Paliouras and Yannis Ioannidis, "OurDMOZ: A System for Personalizing the Web," In Proc. of 6th International Workshop on Personalized Access, Profile Management, and Context Awareness in Databases, 2012.

[10] Dimitrios Pierrakos, Georgios Paliouras, Christos Papatheodorou and Constantine D. Spyropoulos, "Web Usage Mining as a Tool for Personalization: A Survey," ACM, User Modeling and User-Adapted Interaction, Vol.13, No.4, 2003, pp.311-372.

[11] Georgios Paliouras, Christos Papatheodorou, Vangelis Karkaletsis and Constantine D. Spyropoulos, "Clustering the Users of Large Web Sites into Communities," In Proc. of 7th International Conference on Machine Learning, 2000, pp.719-726.

[12] http://wordpress.org/extend/plugins/wp-greet-box/

[13] http://www.adobe.com/products/testandtarget.html

[14] http://www.ariadne.ac.uk/issue28/personalization

[15] http://www.prnewswire.com/news-releases/magiq-the-on-demand-digital-marketing-company-launches-with-website-content-personalization-and-trigger-based-online-marketing-services-for-marketers-70273237.html

[16] http://www.w3c.org/P3P/

[17] http://www-142.ibm.com/software/products/us/en/personalized-product-recommendations/

[18] Ilham Esslimani, Armelle Brun and Anne Boyer, "Densifying a behavioral recommender system by social networks link prediction methods," Springer, Social Network Analysis and Mining, Vol.1, No.3, 2011, pp.159-172.

[19] J. Han and Kamber, "DataMining: Concepts and Techniques," Morgan Kaufmann Publishers.

[20] Jaideep Srivastava, Robert Cooley, Mukund Deshpande and Pang-Ning Tan, "Web usage mining: discovery and applications of usage patterns from Web data," ACM SIGKDD Explorations Newsletter, Vol.1, No.2, 2000, pp.12-23.

[21] Jose M. Domenech and Javier Lorenzo, "A Tool for Web Usage Mining," In Proc. of 8th International Conference on Intelligent Data Engineering and Automated Learning , 2007.

[22] Joseph A. Konstan and John Riedl, "Recommender systems: from algorithms to user experience," Springer, User Modeling and User-Adapted Interaction, Vol.22, No.1-2, pp.101-123.

[23] K. Vinodh and Saji K Mathew, "Web Personalization in Technology Acceptance," in Proc. of IEEE 4th International Conference on Intelligent Human Computer Interaction, 2012, pp.1-6.

[24] Karen H. L. TsoSutter, Leandro Balby Marinho and Lars Schmidt-Thieme, **"Tag-aware recommender systems by fusion of collaborative filtering algorithms,"** In Proc. of 2008 ACM symposium on Applied computing, 2008, pp.1995-1999.

[25] M. D. Mulvenna and A. G. Buchner, "Data mining and electronic commerce," In Proc. of Overcoming Barriers to Electronic Commerce, 1997, pp.1-7.

[26] Magdalini Eirinaki and Michalis Vazirgiannis, "Web Mining for Web Personalization," ACM Transactions on Internet Technology, Vol.3, No.1, 2003, pp.2-38.

[27] Michael Chau and Hsinchun Chen, "A machine learning approach to web page filtering using content and structure analysis," Elsevier, Decision Support Systems, Vol.44, No.2, 2008, pp.482-494.

[28] Murat Ali Bayir, Ismail Hakki Toroslu, Ahmet Cosar and Guven Fidan "Discovering more accurate Frequent Web Usage Patterns," arXiv0804.1409v1, 2008.

[29] Myra Spiliopoulou, Carsten Pohle and Lukas C. Faulstich, "Improving the effectiveness of a web site with Web usage mining," Springer, LNCS, Vol.1836, 2000, pp.142-162.

[30] Oren Etzioni, "The World Wide Web: Quagmire or gold mine," Communications of the ACM, Vol.39, No.11, pp.65-68.

[31] R. Cooley, B. Mobasher and J. Srivastava, "Grouping Web page references into transactions for mining World Wide Web browsing patterns," Technical Report TR 97-021. Dept. of Computer Science, Univ. of Minnesota, Minneapolis, USA, 1997.

[32] R. Kosala and H. Blockeel., "Web mining research: a survey," ACM SIGKDD Explorations Newsletter, Vol.2, No.1, pp.1-15.

[33] Rakesh Agrawal and Ramakrishnan Shrikant, "Fast Algorithms for Mining Association Rules in Large Databases," In Proc. of 20th International Conference on Very Large DataBases, 1994, pp.487-499.

[34] Rana Forsati, Mohammad Reza Meybodi and Afsaneh Rahbar, "An Efficient Algorithm for Web Recommendation Systems," In Proc. of IEEE/ACS International Conference on Computer Systems and Applications, May 2009, pp.579-586.

[35] Ranieri Baraglia, Claudio Lucchese, Salvatore Orlando, Massimo Serrano and Fabrizio Silvestri, "A privacy preserving web recommender system," In Proc. of ACM symposium on Applied computing, 2006, pp.559-563.

[36] Ronaldo Lima Rocha Campos, Rafaela Lunardi Comarella and Ricardo Azambuja Silveira, "Multiagent Based Recommendation System Model for Indexing and Retrieving Learning Objects," Springer, Communications in Computer and Information Science Vol.365, 2013, pp.328-339.

[37] Samira Khonsha and Mohammad Hadi Sadreddini, "New hybrid web personalization framework," In Proc. of IEEE 3rd International Conference on Communication Software and Networks, 2011, pp.86-92.

[38] Sarabjot Singh Anand and Bamshad Mobasher "Intelligent Techniques for Web Personalization," LNCS, Vol.3169, 2005, Springer, pp.1-36.

[39] Tamas Jambor, Jun Wang and Neal Lathia, "Using Control Theory for Stable and Efficient Recommender Systems," In Proc. of The 21st International Conference on World Wide Web, 2012, pp.11-20.

[40] Tom Mitchell, "Machine Learning," McGraw-Hill

[41] U. Manber, A. Patel and J. Robison, "Experience with Personalization on Yahoo!," Communications of the ACM, Vol.43, No.8, August 2000, pp.35-39.

[42] Utpala Niranjan, Dr.V.V.Krishna, Kanduri Srividya and Dr.V.Khanaa, "Developing a Dynamic web recommendation System based on Incremental Data Mining," In Proc. of IEEE 3rd International Conference on Electronics Computer Technology, 2011, pp. 247 - 252.

[43] V.Chitraa and Dr. Antony Selvdoss Davamani, "A Survey on Preprocessing Methods for Web Usage Data," International Journal of Computer Science and Information Security, Vol.7, No.3, 2010, pp.78-83.

[44] Vivek Arvind. B Swaminathan. and J Viswanathan. K. R., "An Improvised Filtering Based Intelligent Recommendation Technique for Web Personalization," In Proc. of Annual IEEE India Conference, 2011, pp. 1194 - 1199.

[45] Xiaoguang Qi and Brian D. Davison, "Web Page Classification: Features and Algorithms," ACM Computing Surveys, Vol.41, No.2, 2009, Article No.12.

[46] Ya-min WANG, Xue-ling HAN and Xiao-wei LIU, "E-commerce Recommendation System Based on CBR and Web Log Mining," In Proc. of IEEE 18th International Conference on Industrial Engineering and Engineering Management, 2011, pp.311-315.

[47] Yu-Hai tao , Tsung-Pei Hong and Yu-Ming Su, "Web usage mining with intentional browsing data," Expert Systems with Applications , Science Direct,2008.

[48] Yu-Hai tao, Tsung-Pei Hong and Yu-Ming Su, "Web usage mining with intentional browsing data," Expert Systems with Applications, Science Direct, 2008.

[49] Zhiwen Yu, Yuichi Nakamura, Seiie Jang, Shoji Kajita, and Kenji Mase, "Ontology-Based Semantic Recommendation for Context-Aware E-Learning," Springer, LNCS, Vol. 4611, 2007, pp.898-907.

[50] Zi-Ke Zhang, Tao Zhou and Yi-Cheng Zhang "Tag-Aware Recommender Systems: A State-of-the-Art Survey," Springer, Journal of Computer Science and Technology, Vol.26, No.5, 2011, pp 767-777.