

# Application of Web Usage Mining Techniques for Web Personalization: A Survey

Tanvi G Pareek<sup>1</sup>, Shireen Saini<sup>2</sup>, Abharika Nayyer<sup>3</sup>

<sup>1,2,3</sup> Student, School of Computer Engineering, VIT University, Vellore

\*\*\*

**Abstract-** *With enhancements in web innovations the World Wide Web is turning into gigantic information or data storage facility at a fast pace. However, the human ability to peruse, get to furthermore, comprehend content does not increment with that pace. Therefore, providing accurate data to users is becoming difficult for the owners of a website. Thus to provide users an experience of personalized web, provision of personalized web services came into existence. Web Usage Mining is one of the very much preferred methodologies in providing web personalization. It is the application of techniques of data mining to extract knowledge from web. In this paper, we will talk about personalization process and application of web usage mining for it. This survey focuses in usage mining area and thus will give us an idea about the applications of web usage mining techniques for web personalization in numerous*

**Keywords—** Web Personalization, Web Usage Mining, Web log files, Web usage mining applications

## 1. INTRODUCTION

Innovative change has prompted a touchy development of recorded data, with the Web being a tremendous storage facility. Here, giving individuals with access to more data isn't the issue; the issue is that an ever increasing number of individuals explore through expansive and confused Web structures, think that it's hard to get to or get the data they need. Personalization can be the answer for this issue; since its goal is to give clients data they need or need, without searching for it unequivocally. Web Usage Mining is one of the very much preferred methodologies in providing web personalization. It is the application of techniques of data mining to extract knowledge from web. It is one of the three categories of mining, others two being Web Content Mining and Web Structure Mining.

The phases involved in Web Usage Mining are:

(A) Data collection: It is the leading step of Web usage mining. It includes extraction of log information from server log documents. Information can be essentially gathered from three sources:

- a. The server side: These logs contain fundamental data e.g.: name and IP of the remote host, date and time of the demand, the demand line precisely as it originated from the customer, and so forth. This data is typically denoted in standard format.
- b. The Proxy Side: To enhance navigation speed through caching, many Internet Service Providers give proxy server services to the customers. The primary distinction with the server side is that intermediary servers gather information of groups of clients getting to groups of web servers.
- c. The Client Side: Access information can be followed additionally on the customer side by utilizing JavaScript or applets, or even altered browsers.

(B) Data Integration: Data recognition is integrating numerous log files into a solitary record.

(C) Data preprocessing: Real world information might be boisterous or conflicting so we need to preprocess them to make them steady and dependable. So preprocessing stage is vital phase of web usage mining.

Personalization means enhancing peoples' results according to their prior searches relevant to their interests without giving much input from the users' side. It is a very common practice now days so that people can enjoy the best user experience with the least amount of effort. These systems are more flexible, reliable, and dynamic to provide personalized results. Web mining is a subset of data mining which helps to identify involved patterns and useful information from the web in order to provide further personalized results for future findings and experiences on the internet. These techniques are used by multiple organizations to extract the above mentioned information basis the interests of their customers or users so that they can promote their business, study the market dynamics and provide personalized advertisements for better profits etc.

## 2. RESEARCH BACKGROUND

### 2.1 Purpose of Research

The purpose of the survey is to get to know about the various applications of web usage mining techniques that have been utilized or that can be utilized for web personalization.

### 2.2 Research Questions

The questions which this survey intended to find answers to were:

1. What is Web Personalization?
2. Which technique is used in providing web personalization?
3. Why web usage mining is preferred for web personalization?
4. In which areas web usage mining helps achieve web personalization?

## 3. LITERATURE SURVEY

### 3.1 Topic-wise Literature Review

Personalization and Web usage mining:

The goal of personalization in view of Web usage mining is to prescribe an arrangement of articles to the current user as determined by usage patterns. This is done by coordinating the current client session with the usage designs found through Web usage mining.

The procedure of Web personalization in light of Web usage mining comprises of the following stages:

- a. Data preparation and transformation
- b. Pattern discovery
- c. Recommendation

The first stage changes unprocessed Web log documents into exchange information which can be then handled by data mining undertakings. Different information mining procedures can be applied to the exchange of data in the next phase for example clustering, association rule mining, and sequential pattern discovery.

The results are then converted into usage profiles which are then used in the final phase of recommendation. The recommendation engine keeps a track of the current user session along with the user patterns that were discovered to give the user a personalized experience.

Process of personalization: Consists of 3 major phases:

- Data collection: Different types of data about the user, website usage and software and hardware is collected in this phase.
- Data preprocessing: The data is prepared and preprocessed.
- Pattern Discovery.

### 3.2 Overall comparative analysis of contents –

**Table 1. Types of Personalization**

Type	Description	Advantage	Disadvantage
Memorization	Cookies used to store user information such as name browsing history etc.	Does not require intelligent learning	Jeopardizes user privacy
Customization	Inputs user's preferences with the help of forms and then customizes the content based on it.	Static-manual or semi-automatic.	Requires significant input
Recommender Systems	Tries recommending hyperlinks automatically that it considers relevant to the user.		Requires data from forms and questionnaires.

### 3.3 Identification of Pattern

Interesting patterns are detected in web usage data by using statistical tools. These usually consist of the following:

- Association rule mining: This helps us to find the correlation between pages that might not be directly connected and reveal associations that are not known between a group of users with the same likes and dislikes.
- Clustering: Use to group items with similar characteristics together.
- Classification: Classes are used to represent the different types and profiles of users studied, hence using this method we segregate the data items by putting them into appropriate classes.
- Sequential pattern discovery: Patterns like web pages accessed immediately are discovered by this method.

## 4. METHODOLOGIES

a. Data Collection: We collect 3 different types of data that are:

- Data about the user: This contains information like name, phone number, demographics, skills, likes and dislikes, goals and aims. To collect this data we can either ask the user directly or the system can implicitly extract this information.
- Data about Website usage: Contains observable data which consists of actions like clicking on links, making online transactions, printing documents etc. This also includes data derived from further processing the above data.
- Data about software and hardware: This information is helpful to produce adaptations.

b. Data pre-processing: This is the most time consuming step taking up to 60-90% of the completion time of a project. This is essentially done to get the data ready for the application of some mining algorithm. Data is logged, then accuracy checks are performed, data is then put together from a variety of sourced, transforming this data into a session file and the it is finally structured according to the requirements.

The steps involved in this process are:

1. Data Cleaning: This step involves removing the irrelevant data and helps in decreasing the size of the data and removing any false associations
2. User Identification: Most important task, as it identifies all the unique users from the data logs.
3. Session Identification: This step gives the frequency of the times a user visits a particular web page
4. Path Completion: Finds out the source of the request and keeps track of pages involved in the path

5. Data Summarization: After all the above processes are completed, the above data is added to a relational database system for subsequent generalization and computations.
- c. Pattern Discovery and Analysis: After the data has been cleaned and the transactions of the user have been identified a pattern is discovered and analyzed. Interesting patterns are detected in web usage data by using statistical tools. These usually consist of the following:
  - Association rule mining: This helps us to find the correlation between pages that might not be directly connected and reveal associations that are not known between a group of users with the same likes and dislikes.
  - Clustering: Use to group items with similar characteristics together.
  - Classification: Classes are used to represent the different types and profiles of users studied, hence using this method we segregate the data items by putting them into appropriate classes.

## 5. EVALUATION CRITERIA

**Sequential Patterns:** They try to recognize which items are succeeded by other sets of items and are influenced by timestamps. Recognizing sequential patterns can be used to predict subsequent visits and developing good and appropriate designs.

**Clustering:** Page clustering finds similarities between different web pages. User clustering is used to find different sets of users which have common browsing patterns. All these information sources are useful for ecommerce applications and to personalize web content.

## 6.RESULTS AND DISCUSSIONS

### 6.1 What is Web Personalization?

To serve specific needs or choices of a user, layout of the content of a website needs to be modified accordingly. This modification according to a personal use is referred to as web personalization. This can be done either by user looking over a menu of accessible choices or by following his or her conduct, (for example, which pages are used by the user and how frequently) on the website.

### 6.2 Which technique is used in providing web personalization?

After various surveys and going through various research papers we concluded that Web Usage Mining is one of the very much preferred methodologies in providing web personalization.

### 6.3 Why web usage mining is preferred for web personalization?

At the point when information mining strategies are connected on web utilization information with a specific reason according to the user needs, it is known as web usage mining. It is a method for gathering and preprocessing web usage information, and afterward building models that speak to the conduct and interests of clients. Such models can naturally be utilized by personalization frameworks for anticipating client's preferences also, thus improving his familiarity while surfing the website.

The process of web usage mining can be depicted using the below figure:

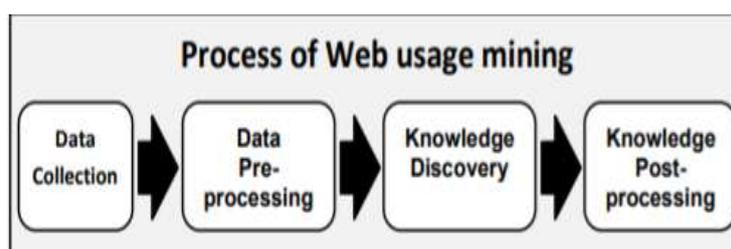


Fig 1. Process of web usage mining

#### 6.4 In which areas web usage mining helps achieve web personalization?

Some of the applications of web usage mining are:

- a. Web personalization
- b. Web design
- c. E-commerce
- d. Web Marketing
- e. Transaction Analysis
- f. Modification of web site
- g. Fraud detection
- h. Customer Relationship Management

After surveying specifically about the applications of web usage mining in web personalization, we got the following results:

- a. Personalization of web usage data mining: This is regarding extracting information based on which website the user visits and how many times does he visit.
- b. Computational Intelligent combinations: Give the distinctive data techniques and systems which have been intended to give Web clients the data they look, before they ask for it explicitly.
- c. Novel online recommender systems: Without the user leading from the front, the system fabricates profiling models and offers recommendations.
- d. Assessing the online users in taking a decision: The objective of a customized site is to exploit the knowledge got from the investigation of the client's navigational conduct in mix with other data gathered, for example, the client's area, past route examples, and things obtained.

#### 7. COMPARATIVE ANALYSIS

The architecture of the system can be viewed like this:

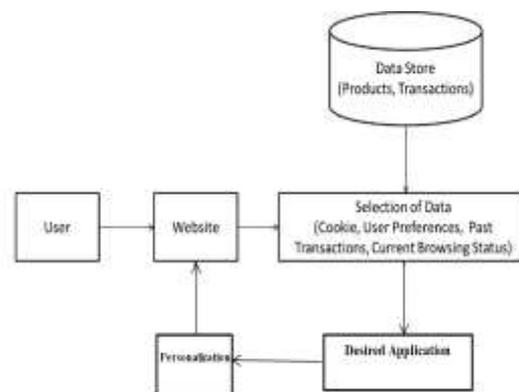


Fig 2. System Architecture

The overall result can be concluded with the help of the following table

**Table- 2. Overall Results**

Parameters	Web Usage Mining
Data View	Interactivity
Main Data	Browser logs Server logs
Representation	Graph Table
Applications	a. Web personalization b. Web design c. E-commerce d. Web Marketing e. Transaction Analysis f. Modification of web site g. Fraud detection h. Customer Relationship Management

## 8. THREATS TO VALIDITY

When a new user comes up on the web, he has no interaction history thus it stands as a problem to the system because there is no existing data to personalize the user's experience. This is called the Called Start or New User Problem as there is no prior information available.

When a new item comes up on the web, it has no ratings as no user has reviewed it before. The systems that depend on item ratings solely (for example, collaborative filtering based approaches) cannot put forward this new item before it has been reviewed a required amount of history has been collected. This is called the Latency Problem or the New Item Problem as again, there is no prior information or ratings for the given item.

More issues also occur while using cookies because the cookie file is stored on the users' computer locally. It gets deleted on closing the session and hence every time a user visits the browser it is considered a new user.

Other issues also arise when user information is gathered through questionnaires or forms, users many a times submit incorrect information and interests which poses as a problem while creating the description and profile of the user.

## 9. CONCLUSION

Thus after reviewing multiple papers we get an insight into the process of personalization of the web and how it is mainly feasible by web usage mining. This paper summarizes the steps that are required to carry out personalization and also categorizes different types of personalization. Web Usage Mining techniques are used to extract browsing patterns and other relevant information basis the interests of their customers or users so that they can be applied to real world problems like promoting their business, study the user dynamics and provide personalized advertisements for better recommendations etc.

## 10. REFERENCES

- [1]. Swami, M., & Kulkarni, M. (2013). Understanding Web Personalization with Web Usage Mining and its Application: Recommender System. *International Journal of Emerging Technology and Advanced Engineering*, 3(5), 726-30
- [2]. Ratnakumar, A. J. (2010). An implementation of web personalization using web mining techniques. *Journal of Theoretical and applied information technology*, 18(1), 67-73.

- 
- [3]. Chaudhary, K., & Gupta, S. K. (2013). Web Usage Mining Tools & Techniques: A Survey. *International Journal of Scientific & Engineering Research*, 4(6), 1762.
- [4]. Singh, S. P. (2016). Web Usage Mining Tools & Techniques: A survey. *International Journal of Advance Research, Ideas and Innovations in Technology*, 2(5).
- [5]. Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. *ACM Sigkdd Explorations Newsletter*, 2(1), 1-15
- [6]. Pierrakos, D., Paliouras, G., Papatheodorou, C., & Spyropoulos, C. D. (2003). Web usage mining as a tool for personalization: A survey. *User modeling and user-adapted interaction*, 13(4), 311-372.
- [7]. Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. N. (2000). Web usage mining: Discovery and applications of usage patterns from web data. *Acm Sigkdd Explorations Newsletter*, 1(2), 12-23.
- [8]. Jagan, S., & Rajagopalan, S. P. (2015). A survey on web personalization of web usage mining. *International Research Journal of Engineering and Technology*, 2(1), 6-12
- [9]. Eirinaki, M., & Vazirgiannis, M. (2003). Web mining for web personalization. *ACM Transactions on Internet Technology (TOIT)*, 3(1), 1-27
- [10]. Brusilovski, P., Kobsa, A., & Nejdl, W. (Eds.). (2007). *The adaptive web: methods and strategies of web personalization* (Vol. 4321). Springer Science & Business Media
- [11]. Mobasher, B., Dai, H., Luo, T., & Nakagawa, M. (2002). Discovery and evaluation of aggregate usage profiles for web personalization. *Data mining and knowledge discovery*, 6(1), 61-