INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN 2278 – 0211 (Online)

Survey Paper on Web Usage Mining for Web Personalization

Namdev Anwat
Department of Computer Engineering
Matoshri College of Engineering & Research Center, Eklahare, Nashik
University of Pune, Pune, India
Varsha Patil
Department of Computer Engineering
Matoshri College of Engineering & Research Center, Eklahare, Nashik
University of Pune, Pune, India

Abstract:
Now a day, World Wide Web (WWW) is a rich and most powerful source of information. Day by day it is becoming more complex and expanding in size to get maximum information details online. However, it is becoming more complex and critical task to retrieve exact information expected by its users. To deal with this problem one more powerful concept is personalization which is becoming more powerful now days. Personalization is a subclass of information filtering system that seek to predict the 'ratings' or 'preferences' that a user would give to an items, they had not yet considered, using a model built from the characteristics of an item (content-based approaches or collaborative filtering approaches). Web mining is an emerging field of data mining used to provide personalization on the web. It consist three major categories i.e. Web Content Mining, Web Usage Mining, and Web Structure Mining. This paper focuses on web usage mining and algorithms used for providing personalization on the web.

Keywords*: Personalization, Web Usage Mining, Data Pre-processing, Pattern Discovery, Pattern Analysis*

1. Introduction

Personalization means persons would get the things or results according to their interests and expectations without giving much more input. Personalization systems are a subclass of information filtering system that seek to predict the 'ratings' or 'preferences' that a user would give to an items, they had not yet considered, using a model built from the characteristics of an item (content-based approaches or collaborative filtering approaches). Personalization systems analyses the individual characteristics and habits without expecting much more input from user and constructs an automated responses to fulfil individual needs. Personalization systems have become very common in recent years. These systems are more flexible, reliable, and dynamic to provide personalized results [1].

Web mining is an emerging field of data mining that automatically extracts useful information and patterns from web documents and constructs personalized results. Web mining techniques are commonly used by various kinds of organizations to extract useful information and patterns on the basis of interest and habits of customers/users. This extracted information is used to promote business, understanding market dynamics, new promotions floating on the internet, personalized advertisements etc [2].

So, in this paper our motivation is to focus on web usage mining, process of usage mining, and algorithms used for usage mining. The rest of the paper is organized as follows: Chapter (2) introduces the related work done on personalization by different researchers using web usage mining. Chapter (3) briefly introduces web mining taxonomy. Chapter (4) gives an overlook of web usage mining, its process and algorithms. Finally, Chapter (5) concludes the discussion of web usage mining.

2. Related Work

Now a day, Personalization systems becoming more popular because of it analyzes individual differences of each user and provides personalized responses according to each individual's needs, interests and preferences. It performs modifications concerning the contents or even the structure of web site dynamically. So, number of researchers has done much on in this area. Some related interesting research efforts are discussed below.

Nasraoui and Krishnapuram et al. [2000] discovered the user session files and formulated groups on the basis of similar characteristics using fuzzy algorithms [3]. According to their research a user or a page can have more than one cluster. In their proposed approach,

after preprocessing of usage data dissimilarity matrix of preprocessed data is created. This is used by fuzzy algorithms in order to cluster typical user session.

Mobasher et. al. [2000] proposed most advanced system, "WebPersonalizer" [4]. It is a powerful framework for mining web log files to extract the useful information for the purpose of recommendations based on the browsing similarities of current user to previous user. After collecting and cleaning of usage data (creating various abstractions of collected data), data mining techniques such as association rule mining, sequential pattern discovery, clustering, and classification are applied in order to discover interesting usage patterns.

The most important contribution of Berendt [2001] in the area of web usage mining is STRATDYN (Strategic and Dynamic) [5] add-on module. It determines the differences between navigational patterns of user and then it exploits the site semantics in the visualization of the results. In this approach, web pages are grouped together on the basis of concept hierarchies. He focused on "interval based coarsening" technique for usage data at different levels of abstraction. For this purpose he used basic and coarsened stratograms for visualization of the results.

Magdalini Eirinaki et. al. [2003] focused on web usage mining. This process relies on the application of statistical and data mining methods to the web log data, resulting in a set of useful patterns that indicate user's navigational behavior [1]. In this approach various data mining algorithms are applied to find navigational patterns.

All above researchers have mostly focused on web usage mining and user profiling for personalized recommendations using search queries and their approach is to provide general personalized environment for all kinds of web users.

3. Web Mining Taxonomy

Web mining technology is emerging field of data mining for WWW based information and resources. The basic focus of web mining is to use data mining techniques and algorithms to extract useful and hidden patterns from unstructured and huge web data or resources. Web mining taxonomy is divided into three categories according to sources of web data [2]. These categories are Web content mining, Web usage mining, and Web structure mining shown in Figure 1. As you can see from this Figure;

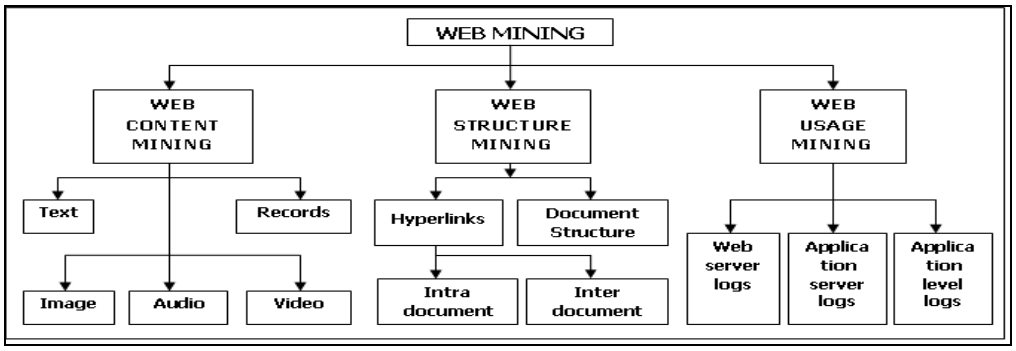

Figure 1: Web mining overview

- **Web content mining** means the extraction of useful information and web knowledge from web sources or web contents such as text, image, audio, video, and structured records [2].
- **Web usage mining** is the application of data mining techniques to find out interesting patterns from web usage data. It mainly tries to extract useful and interesting patterns from usage data such as server logs, client browser logs, proxy server logs, cookies, user sessions, registration data, mouse clicks, user queries, bookmarks etc. and any other data as the results of user interactions [6].
- **Web structure mining** tries to identify the structure of hyperlink in html documents and deduce knowledge [2].

Personalization is a typical application of Web Mining, which can be used to improve web site usage by customizing the contents of a web site with respect to the visitor's need. The personalized web content can take the form of recommended links or items, targeted advertisements, or adjusted text and graphics. The purpose is to customize the interactions on a web site depending on the user's explicit and/or implicit interests, habits, and desires. Web mining technology helps service providers to improve their services by gaining general knowledge about the different user groups. More importantly it can help to provide personalized interfaces and services according to individual characteristics of each user [1]. Figure 2 shows typical web mining process for web personalization.

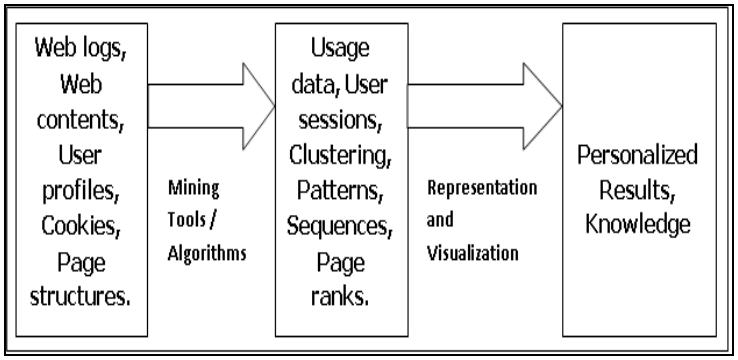

Figure 2: Typical web mining process

4. Web Usage Mining (WUM)

In this article we are interested in web usage mining that uses web data sources in order to discover hidden knowledge about users and their behavior on the Web. Such knowledge, if taken advantage of, brings to organization nothing than benefits and leads directly to profit increase. Site modification, business intelligence, system improvement, personalization and usage characterization are the areas in which the potentials of Web usage mining have been recognized and extensively used.

4.1. Web Usage Mining Process

There are three basic steps that the web usage mining process must follow. These steps are *data preprocessing*, *pattern discovery* and *pattern analysis* [8]. To successfully complete an analysis of a web site, we must obtain data suitable for data mining at the beginning of a process. Most of the authors in their papers agree that data preprocessing step is the most time-consuming step in web usage analysis projects (from 60 to 90 % of the time necessary for the completion of an entire project [8]). The task of data preprocessing is to prepare the data for the application of some data mining algorithm. After data has been preprocessed, it is ready for the application of knowledge extraction algorithms. When exposed to these algorithms, data in web access logs can be transformed into knowledge, most commonly, about association rules, sequential patterns and user clusters. The last phase in WUM process is the analysis of the obtained results in order to distinguish trivial, useless knowledge from knowledge that could be used for Web site modifications, system improvement and/or web personalization. Figure 3 shows the process of web usage mining.

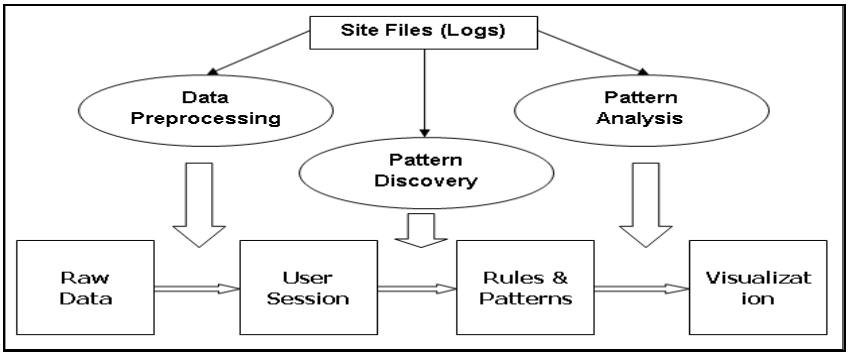

Figure 3: Web usage mining process.

4.1.1. Data Preprocessing

The most important task of the Web Usage Mining process is data preparation. This process is diagrammatically represented in Figure 4. The success of the project is highly correlated to how well the data preparation task is executed. It is of utmost importance to ensure, every nuance of this task is taken care of. This process deals with logging of the data; performing accuracy check; putting the data together from disparate sources; transforming the data into a session file; and finally structuring the data as per the input requirements [7]. The section below describes pre-processing task in detail.

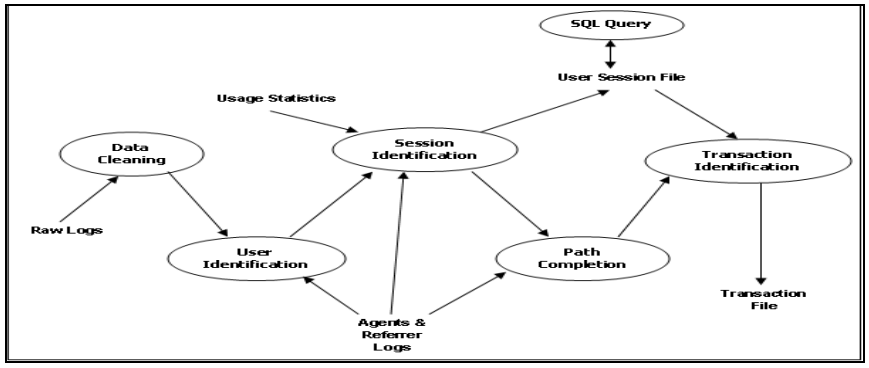

Figure 4: Data pre-processing task

Data Cleaning, The data cleaning process involves removing the irrelevant data from the database log. This data can be in the form of requests from a non-analyzed source, data with missing attributes or the attributes that are not needed for the project goal [9]. This step helps in reducing the size of the data to a great extent. This reduction in size also helps in removing any false associations that could have been created because of this data. When a request to a web page is made, there are various attributes that are called and a lot of contents are loaded in that request. This includes the image files and graphics that are loaded with the web page because of the HTML tags. Since we are interested only in the data that is requested by the user and not any system generated data, we need to make sure that only the user requested data is present in the server logs. Therefore, any of the system-generated data should be avoided and removed from the log files. We can remove these image files, since the kind of website that is being used does not contain a picture archive or any kind of image data. If it does, then it is not recommended to remove these files, as this can be crucial information for determining user behavior. The entries in the log files with the suffix jpg, jpeg, JPEG, class files, .ico files, style sheets and gif files can be removed as these entries do not contribute to the interest of the project. The HTTP error code instances can also be deleted, but can be a useful source of information for network link analysis.

User Identification, As shown in Figure 4, this is the next step after cleaning of data. This is the most important task; one must identify all the unique users from the data logs. Following certain heuristics to identify the users uniquely can accomplish this. Suppose, if the data log shows the users with a common IP address, tracking the agents from which the requests have been made can still differentiate them obtaining different user sessions [8]. With the help of the referrer log from the data, we can generate the different access paths taken by the respective users. One should always keep in mind that these are simple heuristics and cannot always provide the correct information. For example, there can be a scenario where two users have the same common IP address and use the same browser agent for a page request. In this case, 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.

Session Identification, As the name suggests, session identification defines the number of times the user has accessed a web page [8]. We can use the time out mechanism to identify the access time of the user for a respective web page. The time out mechanism basically defines a time limit for the access of a particular page and this limit is usually 30 minutes. Therefore, if the user has accessed the web page for more than 30 minutes, this session will be divided into more than one session. This approach lets us develop some user statistics and helps us in identifying if the user is no longer accessing the requested page. This mechanism faces the problem of caching from browsers and produces incomplete web logs.

Path Completion, This process makes certain, where the request came from and what all pages are involved in the path from the start till the end. The referrer plays an important role in determining the path for a particular request. The problem faced in this process is of the missing entries that mislead in tracking the request [8]. But with the help of the referrer, the site topology and proper tracking of the web page requests, one can easily get the details of the path followed. All of these processes of user identification, session identification and path completion together form the data-structuring phase of the classical data preprocessing scheme. In this process, the preprocessed data is basically formatted according to the needs of the respective data mining algorithms, which are applied to extract important information from the preprocessed data. The formatting of data differs from the kind of algorithms that are used.

Data Summarization, This is one of the advanced data preprocessing tasks that are performed after all of the above processes. In this process, the data is inserted into a relational database system for further generalization and computations. This process of generalization helps in reducing the dimensionality of the data. The aggregated data computation refers to building new parameters from the existing data and adding these parameters, which will help in obtaining new information from the data. One can also perform different analysis mechanisms like the principal component analysis on this data.

After all of the above listed stringent preprocessing techniques, the final data that is produced should be flawless and ready for data mining to contribute in producing correct and effective results.

4.1.2. Pattern Discovery and Analysis

The discovery of user access patterns from the user access logs, referrer logs, user registration logs etc is the main purpose of the Web Usage Mining activity. Pattern discovery is performed only after cleaning the data and after the identification of user transactions and sessions from the access logs. The analysis of the pre-processed data is very beneficial to all the organizations performing different businesses over the web [8]. The tools used for this process use techniques based on AI, data mining algorithms, psychology, and

information theory. The different systems developed for the Web Usage Mining process have introduced different algorithms for finding the maximal forward reference, large reference sequence, which can be used to analyze the traversal path of a user. The different kinds of mining algorithms that can be performed on the preprocessed data include path analysis, association rules, sequential patterns, clustering and classification. It totally depends on the requirement of the analyst to determine which mining techniques to make use of.

Association Rules, This technique is generally applied to a database of transactions consisting of a set of items. This rule implies some kind of association between the transactions in the database. It is important to discover the associations and correlations between these set of transactions. In the web data set, the transaction consists of the number of URL visits by the client, to the web site. It is very important to define the parameter support, while performing the association rule technique on the transactions. This helps in reducing the unnecessary transactions from the database. Support defines the number of occurrences of user transactions within the transaction log. The discovery of such rules from the access log can be of tremendous help in reorganizing the structure of the web site. The frequently accessed web pages should be organized in their order of importance and be easily accessible to the users.

Clustering and Classification, The clustering and classification discovery rules allow grouping the items with similar attributes together. Therefore, when new data is added to the database, it can be classified on the basis of its attributes. In the web transaction data set, the clustering can result in forming clients with similar interests or clients that visit the specific web page based on their demographic information and access patterns. The clustering of clients into specific groups can help in forming business strategies for the future. For example, the organization can develop automated return mail systems for the clients falling in a certain cluster and also develop dynamic changes in the website on the visit from that particular cluster of clients. This can help organizations to become client centric by serving to the interests of their clients and developing a one to one relationship with their clients.

4.2. Web Usage Mining Algorithms

Association rules are a data mining technique that searches for relationships between attributes in large data sets. They can be formally represented as [9]:

$X \rightarrow Y$ (1)

having X, $Y \in D$, D representing the set of all attributes, the so called *itemset* and $X \cap Y = \emptyset$.

If T denotes all transactions t, such that $t \in T$, and if there is an attribute X in transaction t, $X \subset t$, there is probably an attribute Y in t as well, $Y \subset t$ The possibility of this happening is called association rule confidence, denoted by *c* and measured as a percentage of transactions having Y along with X compared to the overall number of transactions containing X. Another important parameter describing the derived association rule is its support, denoted by *s*. It can be calculated as a percentage of transactions containing X and Y to overall number of transactions. These two metrics determine the significance of an association rule. Since the association rules tend to find relationships in large datasets, it would be very time and resource consuming to search for the rules among all data. Because of this each algorithm for discovering association rules begins with the identification of so called *frequent itemsets*. The most popular algorithms use two approaches for determining these itemsets. The first approach is BFS (breath-first search) and is based on knowing all support values of $(k-1)^{th}$ itemset before calculating the support of the k^{th} itemset. DFS (depth-first search) algorithms determine frequent itemsets based on a tree structure [9]. The best known algorithms for mining association rules are Apriori, AprioriTID, STEM, DIC, Partition Algorithm, Elcat, FP-grow, etc.

In web usage mining, association rules are used to discover pages that are visited together quite often. Knowledge of these associations can be used either in marketing and business or as guidelines to web designers for (re)structuring web sites. Transactions for mining association rules differ from those in market basket analysis as they cannot be represented as easily as in MBA (items bought together). Association rules are mined from user sessions containing remote host, user id, and a set of URL's. As a result of mining for association rules we can get, for example, the rule: $X, Y \rightarrow Z$ (c=85%, s=1%). This means that visitors who viewed pages X and Y also viewed page Z in 85 % (confidence) of cases, and that this combination makes up 1% of all transactions in preprocessed logs. In [Cooley et al., 1999] a distinction is made between association rules based on a type of pages appearing in association rules. They identify Auxiliary-Content Transactions and Content-only transactions. The second one is far more meaningful as association rules are found only among pages that contain data important to visitors.

Another interesting application of association rules is the discovery of so called *negative associations*. In mining negative association rules ($X \rightarrow \neg Y$) items that have less than minimum support are not discarded. Algorithms for finding negative association rules can also find indirect associations.

Sequential patterns are another technique for pattern discovery commonly used for discovering knowledge in web access logs. Essentially sequential patterns differ from association rules because they consider the influence of time (timestamp). These timestamps are found in web access logs. Sequential patterns are trying to discover which items are followed by another set of items. For mining sequential patterns from web access logs it is required that each transaction contains the [date] field and a field that denotes the period of time for which we are mining sequential patterns [9]. For example, 10% of visitors who visited page X followed up to page Y. This percentage is called support. Discovering sequential patterns can be used for predicting future visits and developing suitable Web site interface designs for them.

Clustering determines which elements in a dataset are similar. In web usages mining various clustering techniques are applied both for page clustering and user clustering [9]. Page clustering tends to find information about similarities between web pages based upon visits. User clustering tries to discover groups of users having similar browsing patterns. Such knowledge is especially useful in

Ecommerce applications for inferring user demographics in order to perform market segmentation while in the evaluation of Web site quality this knowledge is valuable for providing personalized Web content to the users.

5. Conclusion

This paper gives an insight into the possibility of merging data mining techniques with web access logs analysis for achieving a synergetic effect of web usage mining and its utilization in web personalization. The paper describes the data preprocessing, pattern discovery, and pattern analysis steps, as three basic steps in the process of WUM, which web designers should follow in knowledge extraction. The selection of association rules discovery algorithm as a WUM technique by no means be understood as a suggestion that it is the best WUM algorithm, but as a convenient framework for a research. It is hard, if not impossible, to declare that one data mining algorithm is the best in general, because the possible outcomes of WUM process always depend on the problem in hand. Despite the difference in frameworks, knowledge hidden in click stream data discovered in WUM process, could and should be used for web personalization and further for undertaking corrective actions and making consequent improvements of the previous application design.

6. References

1. Magdalini Eirinaki and Michalis Vazirgiannis, “Web mining for web personalization”, ACM Transactions on Internet Technology, 03(01):1-27, February 2003.
2. Raymond Kosala and Hendrik Blockeel, “Web mining research: A survey”, SIGKDD Explorations, pages 95-104, July 2000.
3. Nasraoui O., Frigui H., Krishnapuram R., and Joshi A, “Extracting web user profiles using relational competitive fuzzy clustering”, IJAI Knowledge Discovery, 09(04):8-14, April 2000.
4. Mobasher B., Cooley R., and Srivastava J, “Automatic personalization based on web usage mining”, ACM Communication, 43(08):142-151, August 2000.
5. Berendt B, “Understanding web usage at different levels of abstraction: Coarsening and visualizing sequences”, ACM SIGKDD Knowledge discovery & Data mining, 04(07):104-108, August 2001.
6. Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pangning Tan, “Web usage mining: Discovery and applications of usage patterns from web data”, ACM SIGKDD Explorations, 01(03):187-192, January 2000.
7. Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava, “Data preparation for mining world wide web browsing patterns”, Knowledge and Information Systems, 01(01):84-89, February 1999.
8. Sasa Bosnjak, Mirjana Maric, Zita Bosnjak, “The role of web usage mining in web application evaluation”, Management Information Systems, Vol. 05(01):31-36, 2010.
9. Borges J. and Levene M, “Data mining of user navigation patterns”, Springer-Verlag, 1836(08):92-111, April 1999.