Volume 4, Issue 8, August 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites

S. Ramkumar*
Research Scholar,
Karpagam University,
Coimbatore, 641021, India

G. Emayavaramban
Research Scholar,
Karpagam University,
Coimbatore, 641021, India

A. Elakkiya
Research Scholar,
Karpagam University,
Coimbatore, 641021, India

Abstract— This is an innovative work for the field of web usage mining. The main feature of our work a complete framework and findings in mining Web usage patterns from Web log files of a real Web site that has all the difficult aspects of real-life Web usage mining, including developing user profiles and external data describing an ontology of the Web content. We are presenting a method for discovering and tracking evolving user profiles. Profiles are also enriched with other domain-specific information facets that give a panoramic view of the discovered mass usage modes. An objective validation plan is also used to assess the quality of the mined profiles, in particular their adaptability in the face of evolving user behaviour.

Keywords— Web mining, Cookies, Session.

I. INTRODUCTION

Web Mining is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World Wide Web. Web usage mining provides the support for the web site design, providing personalization server and other business making decision, etc. In order to better serve for the users, web mining applies the data mining, the artificial intelligence and the chart technology and so on to the web data and traces users' visiting characteristics, and then extracts the users' using pattern[l]. It has quickly become one of the most important areas in Computer and Information Sciences because of its direct applications in e-commerce, CRM, Web analytics, information retrieval and filtering, and Web information systems.

One important research point in web usage mining is the clustering of web users based on their common properties by analysing the characteristics the clusters. One method to cluster web users is to measure similarity of interests between web users access patterns and then cluster them based on the similarities obtained. By mining web users historical access patterns, not only the information about the web is used but also some behavioral characteristics of users could be identified.

The fast pace and large amounts of data available in these online settings have recently made it imperative to use automated data mining or knowledge discovery techniques to discover Web user profiles. Using Web usage mining techniques that can automatically extract frequent access patterns from the history of previous user click streams stored in Web log files. These profiles can later be harnessed toward personalizing the Web site to the user or to support targeted marketing. Although there have been considerable advances in Web.

In this study, we present a complete framework and a summary of our experience in mining Web usage patterns with real-world challenges such as evolving access patterns, dynamic pages, and external data describing an ontology of the Web content and how it relates to the business actors (in the case of the studied Web site, the companies, contractors, consultants, etc., in corrosion). The Web site in this study is a portal that provides access to news, events, resources, company information (such as companies or contractors supplying related products and services), and a library of technical and regulatory documentation related to corrosion and surface treatment. The portal also offers a virtual meeting place between companies or organizations seeking information about other companies or organizations. Without loss of generality, in the rest of this paper, we will refer to all the Web site participants (organizations, contractors, consultants, agencies, corporations, centers, agencies, etc.) simply as companies [1]-[3].

The motive of mining is to find users' access models automatically and quickly from the vast Web log data, such as frequent access paths, frequent access page groups and user clustering. Through web usage mining, the server log, registration information and other relative information left by user access can be mined with the user access mode which will provide foundation for decision making of organizations. This article provides a survey and analysis of current Web usage mining systems and technologies. The profile based data extraction and evaluation is still have the problem in the web usage mining. The goal or objective of web usage mining is to capture, model and analyze the behavioral patterns and profiles of users interacting with a website. The discovered patterns are usually represented as collections of pages, objects or resources that are frequently accessed by group of users with common needs or interest. The goal is obtaining the profile based evaluation from the web logs.

II. LITERATURE SURVEY

Research based on evaluating web mining model has been undertaken in the last decade, some of the prominent studies are given below. J. Srivastava et al discussed that large volumes of data are gathered automatically by Web servers and collected in access log files. Analysis of server access data can provide significant and useful information. Web Usage Mining is the process of applying data mining techniques to the discovery of usage patterns from Web data and is targeted towards applications. It mines the secondary data derived from the interactions of the users during certain period of Web sessions. Web usage mining consists of three phases, namely preprocessing, pattern discovery, and pattern analysis. Given its application potential, Web usage mining has seen a rapid increase in interest, from both the resaerch and practice communities. In this paper, we applied Kohonen's SOM (Self Organizing Map) to pre-processed Web logs of our university Web server logs (http://www.um.ac.ir/) and extract frequent patterns. Result of this paper would be useful for our university Web site owner [4]. Mike Perkowitz et al, developed web sites that improve themselves by learning from user access patterns. Adaptive webs can make popular pages more accessible, highlight interesting links, connect related pages, and cluster similar documents together. An adaptive web can perform these self-improvements autonomously or advise a site's webmaster, summarizing access information and making suggestion. In this paper we define adaptive web sites, explain and formalize several kinds of improvements that an adaptive site can make, and give examples of applying these improvements to existing sites [5]. Mark Levene et al, determine the main activities of web users, known as "surfing", is to follow links. Lengthy navigation often leads to disorientation when users lose track of the context in which they are navigating and are unsure how to proceed in terms of the goal of their original query. Studying navigation patterns of web users is thus important, since it can lead us to a better understanding of the problems users face when they are surfing. We derive Zipf's rank frequency law (i.e. an inverse power law) from an absorbing Markov chain model of surfers' behavior assuming that less probable navigation trails are, on average, longer than more probable ones. In our model the probability of a trail is interpreted as the relevance (or "value") of the trail. We apply our model to two scenarios: in the first the probability of a user terminating the navigation session is independent of the number of links he has followed so far, and in the second the probability of a user terminating the navigation session increases by a constant each time the user follows a link. We analyze these scenarios using two sets of experimental data sets showing that, although the first scenario is only a rough approximation of surfers' behavior, the data is consistent with the second scenario and can thus provide an explanation of surfers' behavior [6]. In this study we present a complete framework for mining Web usage patterns with real-world challenges such as evolving access patterns, dynamic pages, and external data describing ontology of the Web content and how it relates to the business actors. The Web site in this study is a portal that provides access to news, events, resources, company information and a library. The Web site in our study is managed by a nonprofit organization that does not sell anything but only provides free information. Here we perform clustering of the user sessions extracted from the Web logs to partition the users into several homogeneous groups with similar activities and then extract user profiles from each cluster as a set of relevant URLs. Data mining techniques have been applied to extract usage patterns from Web log data, this process is known as Web usage mining.

III. RESEARCH METHODOLOGY

The proposed system generally includes the following several methodologies: data collection, data pre-processing, session based clustering and multi parameter calculation and segmentation.

A. Data collection

Data collection is the first step of web usage mining, the data authenticity and integrality will directly affect the following works smoothly carrying on and the final recommendation of characteristic service's quality. Therefore it must use scientific, reasonable and advanced technology to gather various data. At present, towards web usage mining technology, the main data origin has three kinds: server data, client data and middle data (agent server data and package detecting). The first step in the Web usage mining process consists of collecting the relevant Web data, which will be analyzed to provide useful information about the users' behavior. There are two main sources of data for Web usage mining, corresponding to the two software systems interacting during a Web session: data on the Web server side and data on the client side. In addition, when intermediaries are introduced in the client–server communication, they can also become sources for usage data, like proxy servers and packet sniffers. We will consider each of these sources in the following paper. Also we are trying to associate the data collection methods with the requirements imposed by different classes of personalization functions.

B. Data Preprocessing

Data preprocessing is an important and critical step in the data mining process, and it has a huge impact on the success of a data mining project. The purpose of data preprocessing is to cleanse the dirty/noise data, extract and merge the data from different sources, and then transform and convert the data into a proper format. From the technical point of view, Web usage mining is the application of data mining techniques to usage logs of large data repositories. The purpose of it is to produce result that can be used to improve and optimize the content of a site. In this phase, the starting point and critical point for successful log mining is data extraction. The next task after data extractions are data cleaning and data filtering. Since the origin web logs data sources are blended with irrelevant information, data pre processing acts as an important steps to filter and organize only appropriate information before presenting to any web mining algorithm. An entry of Web server log contains the time stamp of a traversal from a source to a target page, the IP address of the originating host, the type of request (GET and POST) and other data. Many entries that are considered uninteresting for

mining were removed from the data files. The filtering is an application dependent. While in most cases accesses to embedded content such as image and scripts are filtered out. However, before applying data mining algorithm, data preprocessing must be performed to convert the raw data into data abstraction necessary for the further processing.

C. Session Based Clustering

One important research topic in web usage mining is the clustering of web users based on their common properties. Informative knowledge obtained from web user clusters were used for many applications, such as the pre-fetching of pages between web clients and proxies. This proposed session based clustering presents an approach for measuring similarity of interests among web users from their past access behaviors. The similarity measures are based on the user sessions extracted from the user's access logs. A multi-level scheme for clustering a large number of web users is proposed.

Cookies

Cookies can be used to track individual users thus make the sessionizer task easier. However, the use of cookies also raises the concern of privacy thus it requires the cooperation of the users.

Session

All data that needs to be available to the application across different requests within the same session is called session state or session state data. Examples for such data are shopping basket content, intermediary results of database queries, as well as authorization state information. For storing such data between requests, there are in general two possibilities:

- Sending the state information back to the client. With the next request the current state is transmitted to the server again.
- Keeping the necessary data structures on the server. No session data (except the referencing identifier) is transmitted to the client.

Sending the session state data back to the client is generally not recommended. From a security perspective, the main problem is that the session state can easily be manipulated on the client side if it is not protected appropriately. The purpose of session identification is to allow a web application to identify related incoming requests as such.
The following mechanisms for session identification are common:

- Unique identifier in cookie.
- Unique identifier as URL parameter.
- Unique identifier in path portion of URL.

User sessions (identified by means of a cookie-based protocol) are used to build “Session Clusters” eventually leading to a list of suggestions. It finds groups of strongly correlated pages by partitioning the graph according to its connected components. Each component in turn represents a different class, or cluster, of users. The connected components are obtained in an incremental way by using a derivation of the well-known Breadth-First Search (BFS) visit limited to the nodes involved in the request. Basically, we start from the current page identifier and we explore the component to which it belongs. If there are any nodes not considered in the visit a previously connected component has been split and needs to be identified. We simply apply the BFS again, starting from one of the nodes not visited. Furthermore, in order to limit the number of edges of the graph we applied a threshold.

The task of user and session identification is found out the different user sessions from the original web access log. User’s identification is, to identify who access web site and which pages are accessed. The goal of session identification is to divide the page accesses of each user at a time into individual sessions. A session is a series of web pages user browse in a single access. All the session information’s will be grouped by using the session based cluster method.

Sessionization is the process of segmenting the user activity record of each user into sessions, each representing a single visit to the site. The goal of a sessionization heuristic is to reconstruct, from the click stream data, the actual sequence of actions performed by one user during one visit to the site.

D. Multi Parameter Clustering

In the cluster analysis problem one seeks to partition a finite set of objects into disjoint groups (or clusters) such that each group contains relatively similar objects and, relatively dissimilar objects are placed in different groups. In this paper, a multi parametric clustering model for solving cluster analysis problems is presented. It shows how this model can be used to find optimal solutions for certain variations of the clustering problem or, in other cases, for an approximation of the general clustering problem. In the proposed multi parameter methodology clustering the common interest of web users will be grouped

E. Segmentation

Segmentation algorithms divide data into groups, or clusters, of items that have similar properties. User segmentation is the practice of dividing a user base into groups of individuals that are similar in specific ways relevant to marketing, such as age, gender, interests, spending habits, and so on. Here in this system the segmentation has applied for the user browsing behavioural modal. The interest and usage of web data were grouped according to the number users in the area.

Slicing and dicing visitor data provides greater visibility into their behavior patterns. Personalization can be a very powerful segmentation tactic. Many Web site content management systems dynamically display content based on an incoming visitor's identity. A visitor logs in to a Web site, for example, and sees a personalized greeting. This study has been implemented and evaluated in .Net framework by creating a personal website on the network. The flow of the methodology implementation has been described in the following Fig.1 and user authentication is illustrated in the Fig.2

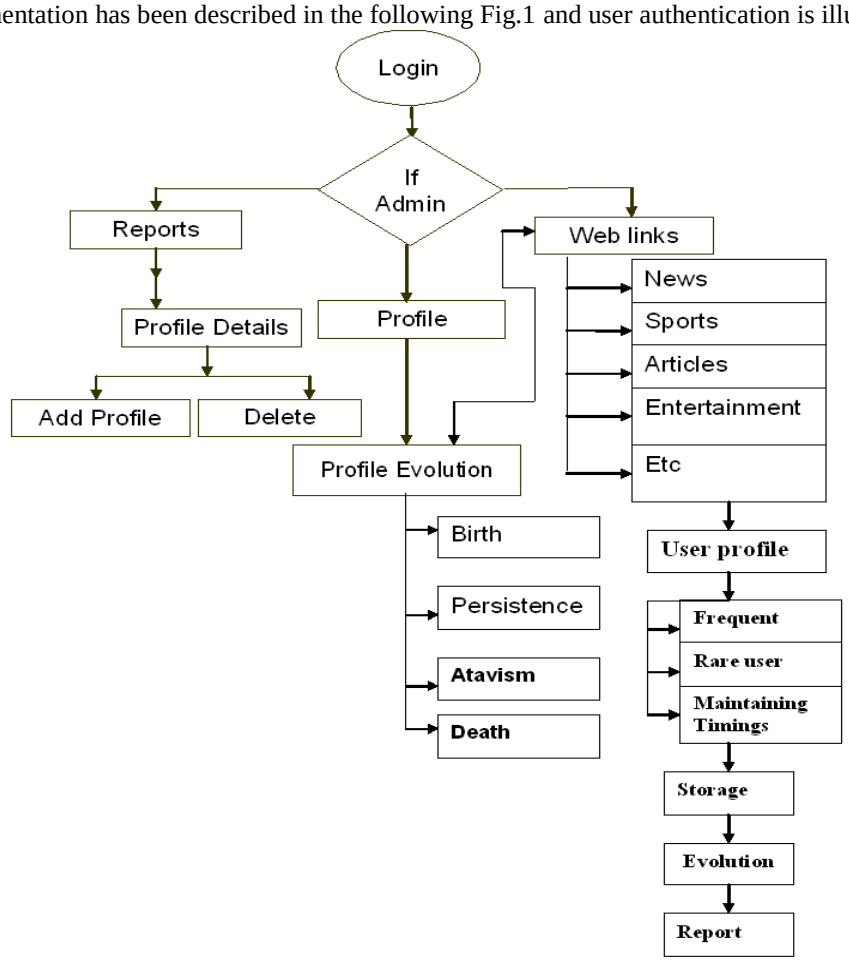

Fig 1. System Design Architecture

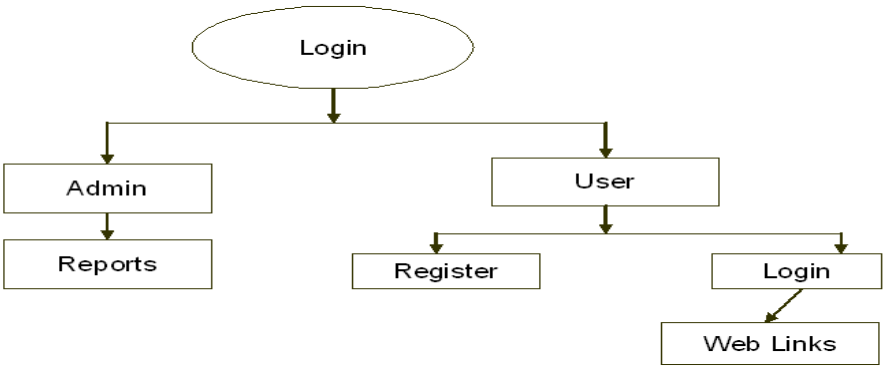

Fig 2. Authentication Architecture

IV. RESULT AND DISCUSSION

Using the methodology and metrics presented above, performed experiments to evaluate the three cluster methods. The results presented in this section provide a detailed picture of the benefits of our approach to personalizing Web directories. The acquired results lead us to the conclusion that although it have obtained good performance by all methods, the use of session based clustering for the personalization of Web directories appears to be the most promising. It helps identifying latent information in the users' choices and derives high-quality community directories that provide significant benefits to their users. The results presented here provide an initial measure of the benefits that we can obtain by personalizing Web directories to the needs and interests of user communities. The main component of a Web personalization system is the usage miner which is shown in Fig.3. Log analysis and Web usage mining is the procedure where the information stored in the Web server logs is processed by applying statistical and data mining techniques such as session based clustering, web log discovery, classification, and sequential multi parameter clustering, in order to reveal useful patterns that can be further analyzed.

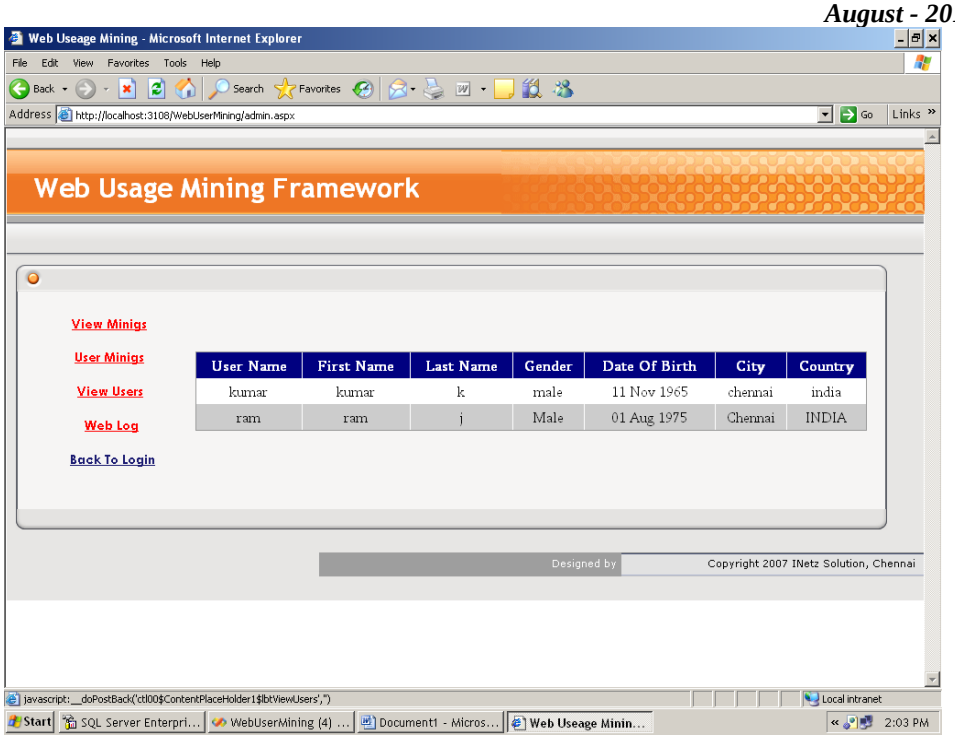

Fig 3.Viewing User Information

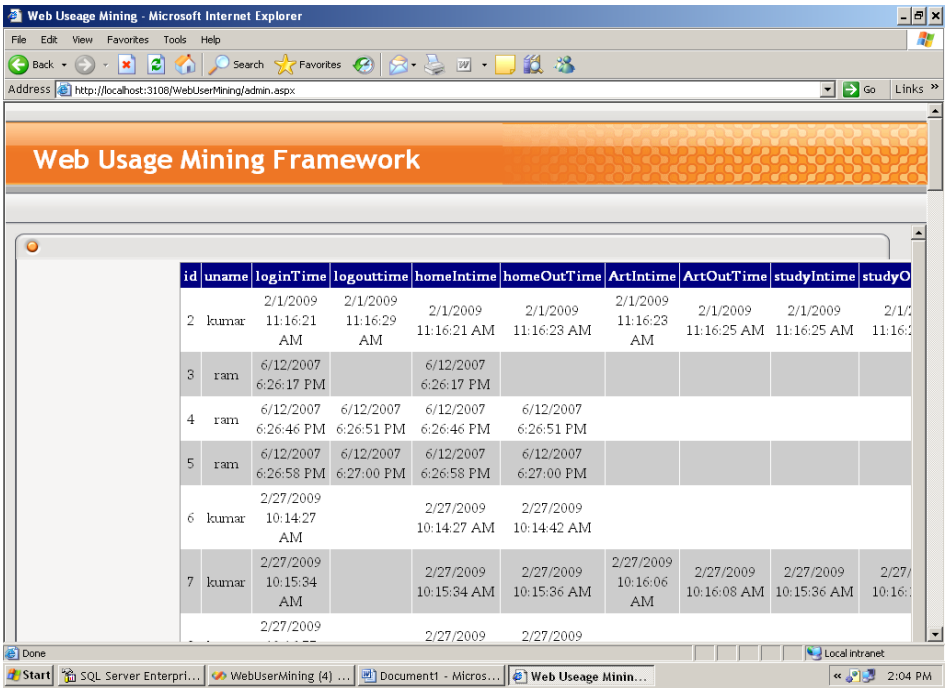

Fig 4. User Audit Information

V. CONCLUSION AND FUTURE WORK

We presented a framework for mining, tracking, and validating evolving multifaceted user profiles on Web sites that have all the challenging aspects of real-life Web usage mining, include evolving user profiles and access patterns, dynamic Web pages, and external data describing ontology of the Web content. A multifaceted user profile summarizes a group of users with similar access activities and consists of their viewed pages, search engine queries, and inquiring and inquired companies. The choice of the period length for analysis depends on the application or can be set, depending on the cross-period validation results. Even though we did not focus on scalability, the latter can be addressed by following an approach similar to, where Web click streams are considered as an evolving data stream, or by mapping some new sessions to persistent profiles and updating these profiles, hence eliminating most sessions from further analysis and focusing the mining on truly new sessions.

In future work, more semantic information will be introduced into mining system so queries of similar meanings can be clustered and generalized. In addition, more log files of longer periods of time (such as months) are required to fabricate more reliable and more useful rules mining algorithm, which will improve further the performance of the web servers.

REFERENCES

[1] R. Cooley, B. Mobasher, and J. Srivastava, "*Web Mining: Information and Pattern Discovery on the World Wide Web*," Proc. Ninth IEEE Int'l Conf. Tools with AI (ICTAI '97), pp. 558-567, 1997.

[2] O. Nasraoui, R. Krishnapuram, and A. Joshi, *"Mining Web Access Logs Using a Relational Clustering Algorithm Based on a Robust Estimator*," Proc. Eighth Int'l World Wide Web Conf. (WWW '99), pp. 40-41, 1999.

[3] O. Nasraoui, R. Krishnapuram, H. Frigui, and A. Joshi, "*Extracting Web User Profiles Using Relational Competitive Fuzzy Clustering*," Int'l J. Artificial Intelligence Tools, vol. 9, no. 4, pp. 509-526, 2000.

[4] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "*Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data*," SIGKDD Explorations, vol. 1, no. 2, pp. 1-12, Jan. 2000.

[5] M. Perkowitz and O. Etzioni, "*Adaptive Web Sites: Automatically Learning for User Access Pattern,"* Proc. Sixth Int'l WWW Conf. (WWW '97), 1997.

[6] Mike Perkowitz and Oren Etzioni*," Data Mining of User Navigation Patterns",* 2009.

