Issues of Learning the Browsing Language

Kavita Das
SoS in Computer Science & IT,
Pt. Ravishankar Shukla University,
Raipur, C.G., India

O. P. Vyas
Prof. & Program Coordinator (s/w Engg.)
IIIT, Allahabad, U.P., India

ABSTRACT
The web is pervading all walks of life and its huge increase in information volume has made the web personalization mandatory. Web Personalization may be achieved by web mining especially the web usage mining technique on the surfing behavior. Learning the surfing behavioral pattern has emerged into a promising research area to achieve web personalization. Till recently web usage mining was done on server logs on website visits and was found insufficient. The goal can be further alleviated if the task or mood of surfer is also learnt. Very shortly, a new approach of web usage mining of client or browser logs on website visits to understand the task or mood of surfer have come into view. It will make the prediction of next web pages by a surfer more accurate. Exploration of browsing behavior at browser to understand an intended task or mood of surfer is hereby termed as 'Browsing Language' as epitome of body language of any person. This work discusses the issues of web usage mining the client side behavior logs for single website only. Implementation of such a capability into a browser will develop the literature of browsing language and may also reduce the overhead of server in realizing web personalization.

General Terms
Web usage mining, web personalization, browsing behavior, browsing task and mood

Keywords: Browsing language

1. INTRODUCTION
Due to continuous proliferation of web, the overhead of surfing has increased in terms of time, effort and semantics. This has lead to high uncertainty in reaching to desired information in due time on the part of surfer. The structure of web is also evolving in order to make electronic coverage of all walks of life. This makes web personalization a highly challenging target in favor of surfers, especially for non-expert surfers. It is suggested that web personalization would require implementation of personalization features in various elements of the web like server, proxy server, and client. Web personalization is a domain that has been recently gaining momentum not only in research area, but also in industrial area. This issue is becoming increasingly important on the Web, as non-expert users are overwhelmed by the quantity of information available online, while commercial Web sites strive to add value to their services in order to create loyal relationships with their visitors-customers.

This study is made to understand the features of browsing behavior of surfers while visiting one particular website and discuss the issues in achieving this target by techniques of data mining. The term 'Browsing Language' is hereby proposed for the complete literature of behavior of surfer to the browser while navigating web. There are multiple factors affecting the browsing activity of a surfer, thus it may be said that the domain of Browser Language is large and complex.

Useful knowledge can be derived by applying statistical techniques and by Web usage mining on the information stored in the web logs. Logs are processed by applying data mining techniques such as association rule discovery, classification, sequential pattern mining and clustering in order to reveal useful patterns concerning the user's navigational behavior, user and page clusters as well as possible correlations between web pages and user groups may be found.

The discovered rules and patterns can then be used for improving the system's performance or for making modifications to the web site. The information included in the web logs can also be integrated with customer profile data, in order to gather prediction intelligence. Web usage mining is a data mining technique, its relation to automated personalization tools is straightforward. It provides rules of predicting the task, intention or mood of surfer from dataset of website usage behavior. The work on Web usage mining can be a source of ideas and solutions towards realizing Web personalization and construct models representing the behavior and the interests of users.

Due to the design web pages, structure of website, goal and intentions of surfer, protocol of client-server client interaction, there are challenges for web usage mining to understand the browser usage behavior of surfer. First, due to continuous interactive usage of browser there is huge amount of data about events, objects and time to be generated. This data requires preprocessing into desirable and relevant format. Second, the mining method of the browsing behavior data needs to be adapted to store possibly large number of complex activity rules and their periodical updating. The adaptation of a website to personalization may simplify the rules in time span. Third, since the behavior of surfer is site specific, the maintenance of the general and the site specific rules of the browsing activities of each website to develop web personalization is long way to go.

The rest of this paper is organized as follows. Section 2 discusses the web personalization's relation to web usage mining. Section 3 describes the research works already done on browsing activity. Section 4 discusses the web usage mining of browsing behavior. Section 5 describes the issues of mining the clients' behavior of browsing one website. In the following sections, this paper is concluded with an outlook into future works.

2. ACHIEVING WEB PERSONALIZATION USING WEB USAGE MINING

2.1 Web Personalization
Web personalization is the adaptability of information systems to the needs of their users. This can be achieved by understanding the surfer's navigational behavior that can be found through the

processing of Web usage logs, as well as the surfer's characteristics and interests. The main component of a Web personalization system is the usage miner. [1, 2, 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 clustering, association rules discovery, classification, and sequential pattern discovery, in order to reveal useful patterns that can be further analyzed.

2.2 Web Usage Mining

Web usage mining refers to the automatic discovery and analysis of patterns in clickstream and associated data collected or generated as a result of user interactions with Web resources on one or more Web sites. The goal is to capture, model, and analyze the behavioral patterns and profiles of users interacting with a Web site. The discovered patterns are navigational patterns that are frequently accessed by groups of surfers with common needs or interests. In Web usage mining, [10] the most basic level of data abstraction is that of a **pageview**. A pageview is an aggregate representation of a collection of Web objects contributing to the display on a user's browser resulting from a single user action. At the user level, the most basic level of behavioral abstraction is that of a **session**. A session is a sequence of page-views by a single user during a single visit. The notion of a session can be further abstracted by selecting a subset of pageviews in the session that are significant or relevant for the analysis tasks at hand. The Web usage mining process consists of gathering 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: the Web server side and the client side and sometimes intermediaries in the client-server communication such as proxy servers and packet sniffers.

Online Analytical Processing (OLAP). OLAP provides a more integrated framework for analysis with a higher degree of flexibility. The data source for OLAP analysis is usually a multidimensional data warehouse which integrates usage, content, and e-commerce data at different levels of aggregation for each dimension.

2.2.1 Server Side Data

The data related to surfers' navigational behavior and web page meta-information while visiting a website, have been generally obtained as server logs of their sessions. [3, 10]The data provides patterns of page visits in the website and show popularity of the pages. The data also suffers from lack of complete information due to various levels of caching of pageviews outside server, lose of data in POST method, and cookies in which contents depend on user cooperation. This data generally provides information about website usage by its various visitors.

2.2.2 Client Side Data

This data may contain the details of pageviews and the interactions of the surfer on the browser at the website. Client side data are more reliable and easy to collect than server side data, since they overcome caching and IP misinterpretation problems, feedback from surfer and the advent of new technology of AJAX [8]. However, the various agents collecting information affect the client's system performance, introducing additional overhead when a user tries to access a Web site, and require the cooperation of users, who may not allow an agent running on their side. This data generally provides information about browser usage by its various visitors. Browser usage may depend on surfer's characteristics such as age and expert-level but may show general patterns w.r.t. a website.

2.3 Web Usage Mining for Web Personalization

Web mining is a complete process. It is broadly divided into three categories, i.e. content mining, structure mining and usage mining. Content mining uses the data consisting of text, graphics, audio, video in the Web pages and finds semantic relationships in them. Structure mining uses inter-page and intra-page link structure and finds popular paths in the links and their weightages. Usage mining uses the data about the pattern of usage of web pages. In the case of Web usage mining this process results in the discovery of knowledge that concerns the behavior of users. Originally, the aim of Web usage mining has been to support the human decision making process and, thus, the outcome of the process is typically a set of data models that reveal implicit knowledge about data items, like Web pages, or products available at a particular Web site. These models are evaluated and exploited by human experts, such as the market analyst who seeks business intelligence, or the site administrator who wants to optimize the structure of the site and enhance the browsing experience of visitors.

The work on Web usage mining can be a source of ideas and solutions towards realizing Web personalization. Usage data, such as those that can be collected when a user browses a specific Web site, represent the interaction between the user and that particular Web site. Web usage mining provides an approach to the collection and preprocessing of those data, and constructs models representing the behavior and the interests of users. These models can be used by a personalization system automatically i.e., without the intervention of any human expert or realizing the required personalization functions. This type of knowledge, i.e., the user models, constitutes operational knowledge for Web personalization. Hence, a Web personalization system can employ Web usage mining methods in order to achieve the required robustness and flexibility. This can be also supported further by orientation of web structure and web content for adapting the usage mining results.

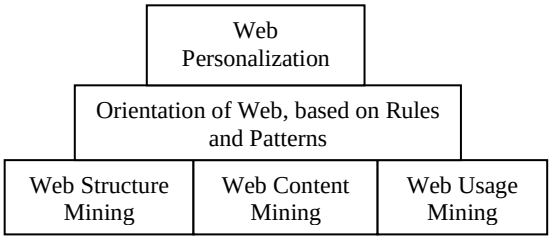

Figure 1. Relation of web mining and web personalization

3. RELATED WORKS

Few works have been contributed towards understanding web browsing activities in [7, 8, 9].

The works [5, 6] has done pioneering research in understanding the tasks and browsing activities of a surfer at the web browser contributed significantly in improving design of the browser. The tasks that were considered significant are Information Gathering, Browsing, Fact Finding and Transaction, and provided their general characteristics visible on the browser.

In [7, 8], the browsing log data on an online newspaper had been collected at an extended browser by few users who have been assigned some predefined tasks related to Information Gathering, Just Browsing, Fact Finding. [8] Data mining techniques were tested on the data. The classification technique C4.5 and association rule mining Apriori had been applied on it. Some classification rules were found with accuracy above 80%. PageViewDuration, TimeSpentOnStartPage, PageviewsPerMinute, were found significant in the rules. Association relationship among some attributes was also found.

[7] Statistical analysis on the same data showed some attributes to be significant in differentiating between the tasks of the surfers. It also found PageViewDuration, TimeSpentOnStartPage, and PageviewsPerMinute to be significant in distinguishing the tasks of the web users.

[9] It made a wide collection of data with fewer restrictions. The detailed log was collected on many websites of the same type i.e. online newspapers and was used to derive behavioral attributes AverageNavigationalDepth, NoOfNewsCategories, TimeSpentOnFrontPage/TaskDuration, and OverAllAreaOfImage to be significant in this job. It also discusses that web site specific attributes are more relevant for predicting the tasks of a browser.

4. THE APPROACH TOWARDS LEARNING BROWSING LANGUAGE

Browsing behavior of the surfer at client end had been researched upon for two purposes – one [4] for designing a more user friendly and efficient browser and the other [7, 8] for understanding the information need of the surfer from his intention using the browsing activities and behavior. This work belongs to the later purpose. It may be considered that learning the behavior of surfer visiting a website by means of observing or logging his way of interaction at browser is similar to understanding the body language of customer visiting a shop. As the interests of the customer can be understood from the visited section of the shop, season/occasion, regular recommendations, queries, history, similar may be our learning strategy at the browser. The main objective is to learn the interests of a surfer in a website so that he can be assisted to reach the needed thing easily and quickly. In other words, the objective is to mould the website into an experienced professional and the browser into an experienced serviceman in the electronic world. The website and the browser should be continuously learning the new situations and refresh the old knowledge periodically. To find the size this period to refresh is again a challenge.

The tasks generally done by a surfer is searching, browsing, fact finding, transaction, information gathering (survey, query, copy&save). During these activities, the surfer interacts with the elements of browser and input devices. The browsing activity can be logged with respect to time, speed, events, section of web page [4,5,6,7,9]. Moreover, there are new challenges. Generally a surfer sticks to a task not more than an upper time limit. Depending on the design of the web page, the surfer may switch between tasks (and mode) before completing one. If the interests of a surfer can be predicted before he switches to other tasks or give restricted option to switch, then his interests can be fulfilled quickly. Generally, the behavior of the surfer dedicated to a task is easy to trace and understand. Thus, surfing can be made an easier task to a non-expert user. The most unpredictable task of a user seems to be the browsing[5,7,8,9]. During this time, a user may switch to any other tasks and return. Thus, understanding the browsing behavior is an important phase to achieve web personalization.

Modeling the browsing behavior of surfer i.e. making the browser understand browsing language of the surfer – is an indicative of possible intentions of user. Thus information requirement of a user may be predicted and supported by other means of web personalization.

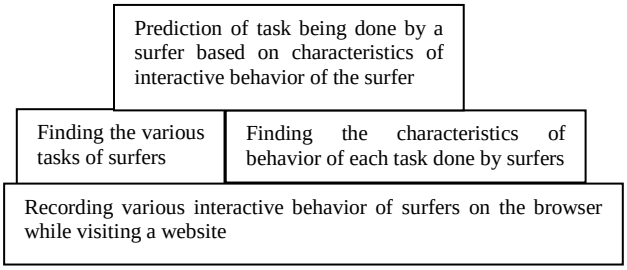

Figure 2. Web Usage mining for browsing language

5. ISSUES OF CLIENT SIDE WEB USAGE MINING

Much research had been contributed on web server logs to understand the usage behavior of surfers using web usage mining. Mining the usage data at client side browsers is a newly developed approach that is showing promising results and also easy to achieve. This work is aims to study the issues to find the behavior of surfer in order to model the personalization capability of a website. It will help to optimize the role of this behavior learning module in the personalization system of a website. It aims to learn the behavior of surfers at browser while visiting only a particular website.

5.1 Attributes and Tasks

The website's structure, depth, content and design of its web pages affect the behavior of a surfer even for the common tasks. This indicates that the browsing characteristics of a task will differ in different websites, at least in different types of websites. For example, online newspaper, online banking system, online electronics shop, online book shop, etc. There are some attributes [ ] that are, in general, significant to distinguish among the tasks of the user. Moreover, the tasks set may not be same for all the websites. Different types of websites may show variant tasks set.

In ebusiness type of websites, where a lot of choices and options are provided, major task may be just browsing with quick switching to and returning from other tasks like Information gathering, transaction, fact finding. Multiple tasks are likely to be overlapping at a time. Hence, the time duration on a page or no. of pageviews may be significantly different from that of online newspaper. In education portals, where the contents are having hierarchical structure with focused subjects, then pure tasks have high probability. Hence, the design affects the values of browsing characteristics of the surfer.

5.2 Data Collection, Preprocessing and Location

The data about various browsing characteristics of surfer will be logged at the browser, which may correspond to only one user or few users. Then there may be 3 options in handling the data periodically-

a. Fusion of the browsing log files from various browsers at the server, find the classification rules then distribute the classification rules to browsers that are visiting the website.
b. Find the classification rules from the log files at a terminal that is visiting the website, then fusion of the rules at the server (with optimization), then distributing the final rules set to the browsers.
c. Find the classification rules from the log files of a terminal that is visiting the website and maintain individual rules set.

These rules set can be used for prediction of future requests from a surfer and caching the relevant pages to provide them to surfers proactively.

In this case, data fusion refers to the merging of log files from several browsers at different terminals accessing the same website.

Data cleaning involves tasks such as, removing extraneous references to embedded objects that may not be important for the purpose of analysis, including references to style files, graphics, or sound files. The cleaning process also may involve the removal of some data fields (e.g. number of bytes transferred or version of HTTP protocol used, etc.) that may not provide useful information in analysis or data mining tasks. Data cleaning also entails the removal of references due to crawler navigations.

5.3 Pageview Identification

Identification of pageviews is heavily dependent on the intra-page structure of the site, as well as on the page contents and the underlying site do-main knowledge. Conceptually, each pageview can be viewed as a collection of Web objects or resources representing a specific "user event," e.g., clicking on a link, viewing a product page, adding a product to the shopping cart. For a static single frame site, each HTML file may have a one-to-one correspondence with a pageview. However, for multi-framed sites, several files make up a given pageview. For dynamic sites, a page-view may represent a combination of static templates and content generated by application servers based on a set of parameters. Logging the pageview at browser event is far more easier and simpler.

5.4 User Identification

Broadly, user identification is not a necessary data required to learn the browsing activities patterns among the users of the particular website. Yet, identification of the terminals' user, place and time accessing the website continuously may reveal few more facts about browsing behavior of the surfers and may refine our concept on browsing language of surfers. It may help to cluster the behavior into number of tasks based on expertise level, age, region, time groups of the surfers. It may provide important knowledge about the clients.

5.5 Sessionization

Sessionization is the process of segmenting the user activity record of each user into sessions, each representing a single visit to the site. The challenge is to mark the start and end points of a real visit, in the scenario where clients can remain connected to the web unlimitedly. In this case, data may be aggregated by predetermined units such as sessions, visitors, or domains.

5.6 Discovery of Patterns and their Updation

The types and levels of analysis, performed on the integrated usage data, depend on the ultimate goals of the analyst and the desired outcomes. Standard statistical techniques can be used on this data to gain knowledge about visitor behavior. This is the approach taken by most commercial tools available for Web log analysis. Reports based on this type of analysis may include in-formation about most frequently accessed pages, average view time of a page, average length of a path through a site, common entry and exit points, and other aggregate measures. Despite a lack of depth in this type of analysis, the resulting knowledge can be potentially useful for improving the system performance, and providing support for marketing decisions. A variety of data mining algorithms are providing in more sophisticated site and customer metrics.

5.7 Visualization

This is a desirable characteristic for visualization of usage rules to facilitate their analysis by the web developers and designers to establish the relationship of the rules and incorporate web personalization at various levels. There is much lack of humanly visualized results generated by data mining algorithms. This makes poorly understandable results whose optimized application becomes a burden.

6. CONCLUSION

This paper explores the various studies towards Web Usage Mining in order to provide more relevant pages to a surfer. It gives the various tasks towards understanding characteristics of surfers that contribute in developing the browsing language literature of web surfers. These characteristics provide useful knowledge about the goals and tasks of a surfer so that relevant pages can be delivered to the surfer.

This study describes the issues important for web usage mining the surfer's interactive behavior at the browser in understanding surfer's behavior for one website only at the client side. It may reduce many overheads from the server machine in fulfilling the web personalization. Moreover much information is easier to obtain about surfers behavior at the client browser. This capability will develop intelligence in browser for learning and predicting the information needs of surfers while visiting a website.

7. OUTLOOK AND FUTURE WORK

Previous works show that website specific attributes give more accuracy in classifying the tasks of any website. It, again, can be proposed that though a set of attributes may be generalized for a particular type of websites yet the classifying values of the attributes will be considerably different depending on the website structure and design.

The tasks of the surfers are also dependent on the type of website, we require an automatic system to derive the tasks and their distinguishing attributes for any website. This is because tasks are dependent on the specific websites.

A general task recognition system may not be very successful in prediction of surfer's task or mood in any website. Websites may be very distinct inviting very different behaviors and tasks from surfers. Such information can be further analyzed in association with the content of a Web site, resulting in improvement of the system performance, users' retention, and/or site modification.

The surfer's pattern of interactive behavior at browser that indicates the task or mood or intention of the surfer at that time period during surfing a website may be termed as 'Browsing Language' in web space as an epitome to 'Body Language' in physical practical world. Thus, it may be said that Browsing Language is having a larger and complex domain than is known yet. An automatic learning system of Browsing Language is required to be explored as the most important component of web personalization system.

8. REFERENCES

[1] Mobasher, B., Cooley, R., and Srivastava, J., 2000, Automatic personalization based on Web usage mining. Communications of the ACM, 2000.

[2] Eirinaki, M., Vazirgiannis, M., 2003, Web Mining for Web Personalization. ACM transactions on Internet Technology, Vol. 3, No. 1, Feb 2003 pp. 1-27

[3] Pierrakos, D., Paliouras, G., Papatheodorou, Spyropoulos C. D., 2003, Web Usage Mining as a tool for Personalization: A Survey. User Modeling and User-Adapted Interaction Volume 13, Number 4, 311-372, 2003

[4] Kellar, M., Watters, C. and Shepherd, M., 2006, The Impact of Task on the Usage of Web Browser Navigation Impact. Proc. of Graphics Interface 2006- portal.acm.org

[5] Kellar, M., Shepherd, M. and Watters, C., 2005, A Field Study Characterizing Web based Information Seeking Tasks. University Ave, Canada 2005.

[6] Kellar, M., Watters, C. and Shepherd, M., 2006, A Goal-based Classification of Web Information Tasks. Proc. of ASIS&T 2006

[7] Gutschmidt, A., Cap, C. H., Nerdinger, F.W. , 2008, Paving the Path to Automatic User Task Identification. Workshop on Common Sense Knowledge and Goal-Oriented Interfaces on the International Conference on Intelligent User Interfaces 2008

[8] Das, K., Vyas, O.P., Cap, C. H. and Gutschmidt, A., 2009, Suitability of Web Usage Mining for Web Content Syndication. Proceedings of National Conference; INDIACom-2009 Computing for Nation Development, Feb 26-27 2009, Bharati Vidyapeeth's Institute of Computer Applications & Management, New Delhi.

[9] Gutschmidth, A., 2010, An approach to situational market segmentation on on-line newspapers based on current tasks. 2010, pp.321-324, Romania, Spain

[10] Cooley, r., Mobasher, B., and Srivastava, J., 1999, Data preparation for mining world wide web browsing patterns. Knowledge and Information Systems, 1(1), 1999

[11] Batista, P., and Silva, M. J., 2002, Mining Online newspaper for Web Access Logs. Proceedings of 2nd International Workshop on Recommendation and Personalization in eCommerce (RPeC'02) (in conjunction with AH 2002)

[12] Sumathi, C. P., Valli, R. P., and Santhanam, T., 2010, Automatic Recommendation of Web Pages in Web Usage Mining. International Journal on Computer science and Engineering (IJCSE) Vol. 02, No. 09, 2010, 3046-3052