Web Personalization Using Web Mining: Concept and Research Issue

Pooja Mehtaa, Brinda Parekh, Kirit Modi, and Paresh Solanki

***Abstract*—Web mining is the application of the data mining which is useful to extract the knowledge. Web mining has been explored to different techniques have been proposed for the variety of the application. Most research on Web mining has been from a 'data- centric' or information based point of view. Web usage mining, Web structure mining and Web content mining are the types of Web mining. Web usage mining is used to mining the data from the web server log files. Web Personalization is one of the areas of the Web usage mining that can be defined as delivery of content tailored to a particular user or as personalization requires implicitly or explicitly collecting visitor information and leveraging that knowledge in your content delivery framework to manipulate what information you present to your users and how you present it. In this paper, we have focused on various Web personalization categories and their research issues.**

***Index Terms*—Web mining, web usage mining, web personalization.**

I. Introduction

Data Mining (the analysis step of the Knowledge Discovery in Databases process, or KDD), a relatively young and interdisciplinary field of computer science, is the process of discovering new patterns from large data sets involving methods from statistics and artificial intelligence but also database management. Web mining is the application of data mining techniques to extract knowledge from Web data including web documents, hyperlinks between documents, usage logs of web sites, etc [1].

Two different approaches were taken in initially defining Web mining. First was a 'process- centric view', which defined Web mining as a sequence of tasks [2]. Second was a 'data- centric view', which defined Web mining in terms of the types of Web data that was being used in the mining process [3].

We decompose the web mining to following sub tasks [4]:

- Resource Finding: the task of retrieving indented Web documents.
- Information Selection and Pre-processing: automatically selecting and pre-processing specific information from retrieved web resources.
- Generalization: automatically discovers general patterns at individual web sites as across multiple sites.
- Analysis: validation and /or interpretation of the mined patterns.

Manuscript received June 18, 2012; revised August 12, 2012.

Pooja Mehtaa, Brinda Parekh, Kirit Modi, and Paresh Solanki are with Department of Information Technology, U. V. Patel College of Engineering, Gujarat, India (e-mail: poojamehta810@gmail.com, brin.prkh@gmail.com).

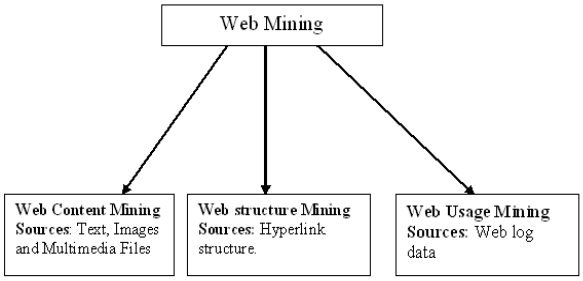

Fig. 1. The types and sources of Web mining

The above Fig. 1 shows the types and sources of Web mining. Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables [5]. Research in web content mining encompasses resource discovery from the web, document categorization and clustering, and information extraction from web pages [6]. Web structure mining studies the web's hyperlink structure. It usually involves analysis of the in-links and out-links of a web page, and it has been used for search engine result ranking. [6]. Web Structure Mining can be regarded as the process of discovering structure information from the Web. This type of mining can be performed either at the (intra-page) document level or at the (inter-page) hyperlink level [5]. Web structure mining is the process of inferring knowledge from the World Wide Web organization and links between references and referents in the Web [7].

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web based applications. It also called as Web log mining. Some of the typical usage data collected at a Web site includes IP addresses, page references, and access time of the users. [5]

Area of Web Usage Mining:

- Personalization
- System Improvement
- Site Modification
- Business Intelligent
- Usage Characterization

The remainder of the paper is structured as follows. First section reviews Web mining and its type. In other section, describes Personalization and its categories and research issues and conclusion are described.

II. Overview of the Various Personalization Categories

Web personalization is a strategy, a marketing tool, and

DOI: 10.7763/IJIET.2012.V2.192

an art. The objective of a Web personalization system is to "provide users with the information they want or need, without expecting from them to ask for it explicitly" [8]. Personalization requires implicitly or explicitly collecting visitor information and leveraging that knowledge in your content delivery framework to manipulate what information you present to your users and how you present it. A personalization mechanism is based on explicit preference declarations by the user and on an iterative process of monitoring the user navigation, collecting its requests of ontological objects and storing them in its profile in order to deliver personalized content [9].

A. Phases of Web Personalization

The Web Personalization process divides in to four distinct phases [5].

- *Collection of Web data*–In this, implicit data includes past activities/click streams as recorded in Web server logs and/or via cookies or session tracking modules. Explicit data usually comes from registration forms and rating questionnaires. In some cases, Web content, structure, and application data can be added as additional sources of data, to shed more light on the next stages.
- *Preprocessing of Web data*–In this, Data is frequently pre-processed to put it into a format that is compatible with the analysis technique to be used in the next step. Preprocessing may include cleaning data of inconsistencies, filtering out irrelevant information according to the goal of analysis. Most importantly, unique sessions need to be identified from the different requests, based on a heuristic, such as requests originating from an identical IP address within a given time period.
- *Analysis of Web Data*–Also known as Web Usage Mining, this step applies machine learning or data mining techniques to discover interesting usage pattern and statistical correlation between web pages and user groups. This step frequently results in automatic user profiling, and is typically applied offline, so that it does not add a burden on the web server.
- *Decision making/Final Recommendation*–It makes use of the results of the previous analysis step to deliver recommendations to the user. It involves generating dynamic Web content on the fly, such as adding hyperlinks to the last web page requested by the user. This can be accomplished using a variety of Web technology options such as CGI programming.

B. Personalization Categories

Here the table shows the details of personalization categories and its description.

TABLE I: PERSONALIZATION CATEGORIES.

Sr. No.	Strategies	Description
1.	Memorization[5]	• Simplest. • Most widespread form of personalization, user information such as name and browsing history is stored (e.g. using *cookies*), to be later used to recognize and greet the returning user. • Implemented on the Web server. • Can also jeopardize user privacy.
2.	Customization[5]	• Takes as input a user's preferences from registration forms in order to customize the content and structure of a web page. • process tends to be static and manual or at best semi-automatic • Implemented on the Web server. • E.g. My Yahoo and Google
3.	Guidance or Recommender Systems[5]	• Tries to *automatically* recommend hyperlinks that are deemed to be relevant to the user's interests, in order to facilitate access to the needed information on a large website. • It is implemented on the Web server. • Relies on data that reflects the user's interest ***implicitly*** (browsing history as recorded in Web server logs) or ***explicitly*** (*user profile* as entered through a registration form or questionnaire).
4.	Task Performance Support[5]	• A personal assistant executes actions on behalf of the user, in order to facilitate access to relevant information. • This approach requires heavy involvement on the part of the user, including access, installation, and maintenance of the personal assistant software. • It also has very limited scope in the sense that it cannot use information about other users with similar interests.
5	Web usage data mining personalization[9]	• The customer preference and the product association are automatically learned from click stream. • In order to avoid the poor recommendations that will lead to disappoint customers, customers who are likely to buy recommended products are selected using decision tree induction.

6	Computational Intelligent combinations[9]	• Provide the different information system which have been designed to provide Web users with the information they search, without expecting them to ask for it explicitly
7	Novel online recommender system[9]	• It builds profiling models and offers suggestions without the user taking the lead.
8	Helping Online Customers Decide through Web Personalization[9]	• The goal of a personalized website is to take advantage of the knowledge obtained from the analysis of the user's navigational behavior in combination with other information collected, such as the user's location, previous navigation patterns, and items purchased.
9	Automatic Personalization Based on Web Usage Mining.[10]	• In which the user preference is automatically learned from Web usage data, by using data mining techniques.
10	Caching[11]	• Efficiently delivering web content, i.e., caching and prefetching. Caching refers to the practice of saving content in memory in the hope that another user will request the same content in near future, while involves guessing at which content will be of interest to the user, and loading it into memory.

III. Reserch issue

The evolution of the algorithms to take into account additional implicit user feedback of the final products chosen and not only the e-shops and services [2].

Researchers must perform more empirical studies that cover different types of online service providers (for example, game malls). Such work will help identify the generic architecture and common components (for example, profiling and matching) reusable in Web personalization [8].

It will be interesting to compare personalized recommender system with a standard collaborative filtering based methodology in the aspect of recommendation performance. And it will also be an interesting research area to conduct a real marketing campaign to customers using this methodology and to evaluate the performance [12].

Extracting useful patterns and rules using data mining techniques in order to understand the users' navigational behavior, so that decisions concerning site restructuring or modification can then be made by humans [5].

Computing similarity between domain objects and aggregate domain-level patterns, as well as learning techniques to automatically determine appropriate combination functions used in the aggregation process. [13]

Can exploit and enable a more effective integration and mining of content, usage, and structure data from different sources promise to lead to the next generation of intelligent Web applications [1].

IV. Conclusion

In this paper, first we have mainly focused on the web mining types- Web content mining, web structure mining and web usage mining. After that, we have introduced the web mining techniques in the area of the Web personalization.

Personalization requires the different goals and also it is useful to develop different business application. E-commerce is one of the example of this personalization technique which depend on the how well the site owners understood the user's behavior and their needs. Web usage mining is useful for the pattern matching, site reorganization, product/site recommendation etc. Future efforts, investigating architectures and algorithms that can exploit and enable a more effective integration and mining of content, usage, and structure data from different sources promise to lead to the next generation of intelligent Web applications.

References

[1] J. Srivastva, P. Desikan, and V. Kumar, *Web mining – Concepts, Application and Research direction*, pp. 51, 2009.

[2] O. Etzioni, "The World-Wide Web, Quagmire or Gold Mine?" *Communications of the ACM*, vol. 39, no. 11, pp. 65–68, 1996.

[3] R. Cooley, J. Srivastava, and B. Mobasher, "Web mining: Information and pattern discovery on the World Wide Web". in *Proc. of the 9th IEEE International Conference on Tools with Artificial Intelligence(ICTAI'97)*, 1997.

[4] R. Kosala and H. Blockeel, "Web Mining Research: A Survey," *SSIGKDD Explorations,ACM SIGKDD*, July 2000.

[5] A. J. Ratnakumar, "An Implementation of Web Personalization Using Web Mining Techniques," J*ournal of Theoretical and applied information technology*, 2005.

[6] W. Bin and L. Zhijing, "Web Mining Research," in *Proceedings of the fifth International Conference on Intelligence and Multimedia Applications (ICCIMA'03)*, 2003.

[7] Q. Han, X. Gao, and W. Wu, *Study on Web Mining Algorithm Based on Usage Mining*, 2010.

[8] M. Eirinaki and M. Vazirgiannis Athens University of Economics and Business, "Web Mining for Web personalization," *ACM Transactions on Internet Technology*, 2005.

[9] D. Antoniou, M. Paschou, E. Sourla, and A. Tsakalidis, "A Semantic Web Personalizing Technique The case of bursts in web visits," presented at IEEE Fourth International Conference on Semantic Computing, 2010.

[10] B. Mobasher, R. Cooley, and J. Srivastava, "Automatic Personalization Based on Web Usage mining-Communication," *ACM*, 2000.

[11] J. Wang, "A Survey of Web Caching scheme for the Internet," *ACM SIGCOMM computer Communication*, 1999.

[12] K. R. Suneetha and R. Krishnamoorthi, "Identifying User Behavior by Analyzing Web Server Access Log File," *IJCSNS International Journal of Computer Science and Network Security*, vol. 9, no. 4, April, 2009.

[13] J. Srivastava, R. Cooley, M. Deshpande, and T. P- Ning, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," *SIGKDD Explorations*, vol. 1, issue 2, pp. 12-23, 2000.