Overview on Web Mining and Different Technique for Web Personalisation

***Pradnyesh Bhisikar [1],Prof. Amit Sahu [2]**

[1]M.E (Scholar), G.H.Raisoni College of Engineering and Management, Amravati.
[2]M.E (CSE),G.H.Raisoni College of Engineering and Management, Amravati.

Abstract

Web mining is mainly focused on learning about web user with their interaction with web sites and application of web to extract knowledge from World Wide Web i.e. WWW. The motive of web mining is to find user's access object automatically and promptly from the huge web log data such as frequent access paths, frequent access groups and clustering of data. This article provides a survey and analysis of current web mining system and technologies. There are three broad category in web mining i) web usage mining ii) web content mining, iii) web structure mining. Which is shown in figure 1. Through web usage mining, performs six major task data gathering, data preparation, navigation pattern discovery, patter analysis, pattern visualisation and pattern application. Through web content mining extract useful information from the contents of web documents. In this paper we implement how Web mining techniques can be apply for the Customization i.e. Web personalization.

Keywords - Navigation Patterns, Pattern Analysis; Content Mining; Structure Mining; User/Session identification; Web Recommender; Web log

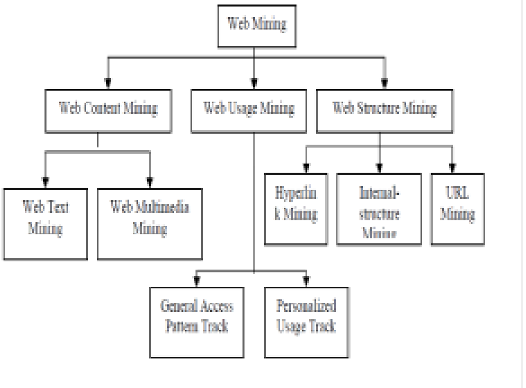

Figure 1 : Classification of web mining

I. INTRODUCTION

Web Mining is the extraction of interesting and potentially useful patterns and implicit information from activity related to the World Wide Web. With the dramatically quick and explosive growth of information available over the Internet, World Wide Web has become a powerful platform to store, disseminate and retrieve information as well as mine useful knowledge. Web usage mining provides the support for the web site design, providing personalization server and other business making decision, etc. Due to the properties of the huge, diverse, dynamic and unstructured nature of Web data, Web data research has encountered a lot of challenges, such as scalability, multimedia and temporal issues etc. A user interacts with the Web; there is a wide diversity of user's navigational preference, which results in needing different contents and presentations of information.

II. WEB MINING TECHNIQUES

Web Content 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.

Graph base web Mining: The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages.

Utilisation in web Mining: Web Utilised Mining is the application of database mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web based applications.

Text Mining**:** Due to the continuous growth of the volumes of text data, automatic extraction of implicit previously unknown and potentially useful information becomes more necessary to properly utilize this vast source of knowledge. Text mining, therefore, corresponds to extension of the data mining approach to textual data and its concerned with various tasks, such as extraction of information implicitly contained in collection of documents or similarity- based structuring.

III. Web Usage Mining

Concept of web usage mining

A] Data accumulation:

Data *accumulation* 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.

B] Data preprocessing:

Some databases are insufficient, inconsistent. The data pre-treatment is to carry on a unification transformation to those databases. The result is that the database will to become integrate and consistent, thus establish the database which may mine. In the data pre-treatment work, mainly include data cleaning, user identification, session identification and path completion.

1) Data Cleaning:

The purpose of data cleaning is to eliminate irrelevant items, and these kinds of techniques are of importance for any type of web log analysis not only data mining. According to the purposes of different mining applications, irrelevant records in web access log will be eliminated during data cleaning.

1. The records of graphics, videos and the format information. The records have filename suffixes of GIF, JPEG, CSS, and so on, which can found in the URI field of the every record.

2. The records with the failed HTTP status code. By examining the Status field of every record in the web access log, the records with status codes over 299 or fewer than 200 are removed. It should be pointed out that different from most other researches, records having value of POST or HEAD in the Method field are reserved in present study for acquiring more accurate referrer information.

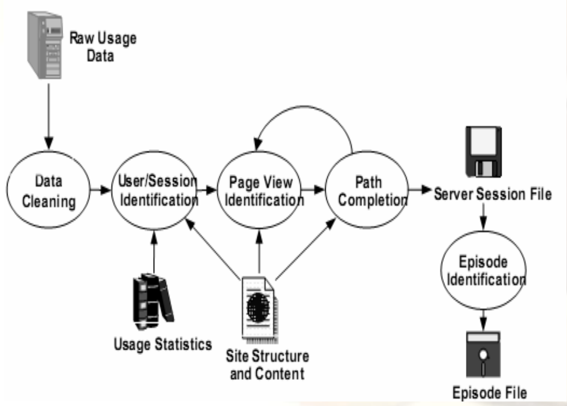

Figure 2: Preprocessing of web usage data

***2*) User and Session Identification:**

The task of user and session identification is find 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

3) Path completion

Another critical step in data pre-processing is path completion. There are some reasons that result in path's incompletion, for instance, local cache, agent cache, "post" technique and browser's "back" button can result in some important accesses not recorded in the access log file, and the number of Uniform Resource Locators(URL) recorded in log may be less than the real one. As a result, the user access paths are incompletely preserved in the web access log. To discover user's travel pattern, the missing pages in the user access path should be appended. The purpose of the path completion is to accomplish this task. The better results of data pre-processing, we will improve the mined patterns' quality and save algorithm's running time.

C] Knowledge Discovery

Use statistical method to carry on the analysis and mine the pretreated data. We may discover the user or the user community's interests then construct interest model. At present the usually used machine learning methods mainly have clustering, classifying, the relation discovery and the order model discovery.

D] Pattern analysis

Challenges of Pattern Analysis are to filter uninteresting information and to visualize and interpret the interesting patterns to the user. First delete the less significance rules or models from the interested model storehouse; Next use technology of OLAP and so on to carry on the comprehensive mining and analysis; Once more, let discovered data or knowledge be visible; Finally, provide the characteristic service to the electronic commerce website.

IV. PERSONALIZATON ON THE WEB MINING

Web personalization is a strategy of marketing tool, and largely art not a science. 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. It is important to detail what it is you hope to do and, from that knowledge, develop an understanding of how you get from an idea to implementation. Web personalization can be seen as an interdisciplinary field that includes several research domains from user modeling [1], social networks [2], web data mining [3,4,2], human-machine interactions to Web usage mining[4]; Web usage mining is an example of approach to extract log files containing information on user navigation in order to classify users. Other techniques of information retrieval are based on documents categories' selection [4]. Contextual information extraction on the user and/or materials (for adaptation systems) is a technique fairly used also include, in addition to user contextual information, contextual information of real-time interactions with the Web. [5] proposed a multi-agent system based on three layers: a user layer containing users' profiles and a personalization module, an information layer and an intermediate layer. They perform an information filtering process that reorganizes Web documents. [6] Propose reformulation query by adding implicit user information. Requests can also be enriched with predefined terms derived from user's profile [5] develop a similar approach based on user categories and profiles inference. User profiles can be also used to enrich queries and to sort results at the user interface level [7]. Other approaches also consider social-based filtering [8] and collaborative filtering. For example, user queries can be enriched by adding new properties from the available domain ontology [8]. User modelling by ontology can be coupled with dynamic update of user profile using results of information-filtering and Web usage mining techniques.

V. Result and Discussion

In this paper we studied on classification on web mining and concatenation with the web personalisation and in that various method. Through web usage mining, performs six major task data gathering, data preparation, navigation pattern discovery, patter analysis, pattern visualisation and pattern application. In the web mining there are we classified in terms web content mining, graph base web mining, utilisation in web mining and text mining.

VI. Conclusion

In this article, we have outlined three different modes of web mining, namely web content mining, web structure mining and web usage mining Web usage mining model is a kind of mining to server logs. Web Usage Mining plays an important role in realizing enhancing the usability of the website design, the improvement of customers' relations and improving the requirement of system performance and so on. We have presented in this paper the significance of introducing the web mining techniques in the area of web personalization. Content, usage, and structure data from different sources promise to lead to the next generation of intelligent Web applications. Use of artificial intelligence in these techniques should be next topic in this area of research.

REFRENCES

[1] Eirinaki M., Vazirgiannis M. (2003). Web mining for web personalization. *ACM Transactions On Internet Technology (TOIT)*, 3(1), 1-27.

[2] Agrawal R. and Srikant R. (2000). Privacy preserving data mining, In Proc. of the ACM SIGMOD Conference on Management of Data, Dallas, Texas, 439-450.

[3] Berendt B., Bamshad M, Spiliopoulou M., and Wiltshire J. (2001). Measuring the accuracy of sessionizers for web usage analysis, In *Workshop on Web Mining*, at the First SIAM International Conference on Data Mining, 7-14.

[4] Mobasher, B., Web Usage Mining and Personalization, in Practical Handbook of Internet Computing, M.P. Singh, Editor. 2004, CRC Press. p. 15.1-37.

[5] Maier T. (2004). A Formal Model of the ETL Process for OLAP-Based Web Usage Analysis. In *Proc. of "WebKDD- 2004 workshop on Web Mining and WebUsage Analysis"*, part of the ACM KDD: Knowledge Discovery and Data Mining Conference, Seattle, WA.

[6] Pierrakos, D., et al. *Web* Community Directories: A New Approach to Web Personalization. in Proceeding of the 1st European Web Mining Forum (EWMF'03). 2003, p. 113-129, Cavtat-Dubrovnik, Croatia.

[7] Kargupta H., Datta S., Wang Q., and Sivakumar K. (2003). On the Privacy Preserving Properties of Random Data Perturbation Techniques, In *Proc. of the 3rd ICDM IEEE International Conference on Data Mining (ICDM'03),* Melbourne, FL.

[8] Linden G., Smith B., and York J. (2003). *Amazon.com* Recommendations Itemto-item collaborative filtering, *IEEE Internet Computing*, 7(1), 76-80.

[9] Schafer J.B., Konstan J., and Reidel J. (1999). Recommender Systems in ECommerce, In *Proc. ACM Conf. E-commerce*, 158-166.