ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

Web Personalization Using Web Usage Mining

C.Nalini, A.Sangeetha, Sundararajan.M, Arulselvi S

Professor, Dept. of CSE, Bharath University, Chennai, Tamil Nadu, India

Ph.D Scholor, Dept. of CSE, Bharath University, Chennai, Tamil Nadu, IndiaDirector, Research Center for Computing and Communication, Bharath University, Chennai, Tamil Nadu, India

Co-Director, Research Center for Computing and Communication, Bharath University, Tamil Nadu, India

ABSTRACT: Web mining is the application of the data mining which is useful to extract the knowledge. 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 in 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 to a particular user or as personalization requires implicitly or explicitly collecting information of the user. 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.

KEYWORDS: Web mining, web usage mining, web personalization**.**

I. INTRODUCTION

The Web can be characterized by the enormous volume and coverage of Web content, the phenomenal number of Web users and businesses, the vast number of computers and devices accessing Web, and the large number of Web-based applications. A survey conducted by OCLC in 2002 revealed that there were 3 million public Websites and 1.4 billion Web pages at that point in time[1]**.** . Users today perform more searches using Web search engines .The ubiquity of Web offers some obvious explanations, namely:
the ability to browse Web content directly on the users’ computers and the ease of downloading them is clearly a big draw; and the availability of Web search engine (e.g., Google2) and Web directories (e.g., Yahoo!3, DMOZ4) has helped tremendously simplified the process of searching Web content. Nevertheless, Web content is not always easy to use. Due to the unstructured and semi-structured nature of Web pages.

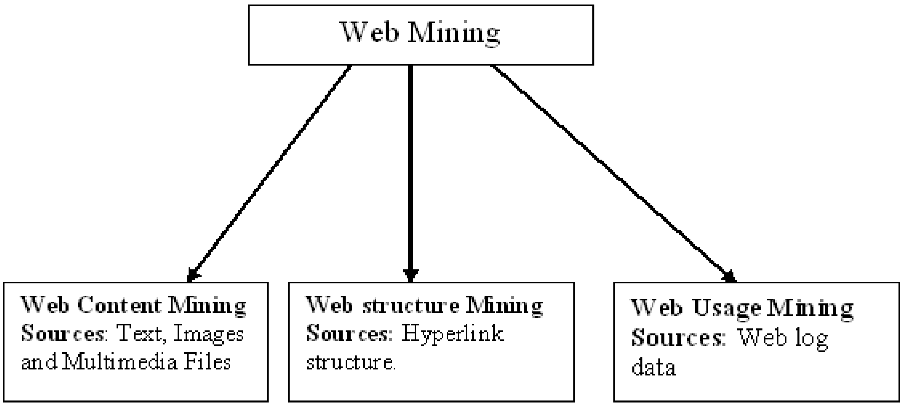

Web Mining
Web Content Mining Sources: Text, Images and Multimedia Files
Web structure Mining Sources: Hyperlink structure.
Web Usage Mining Sources: Web log data

Fig. 1. The types and sources of Web mining

DOI: 10.15680/ijircce.2015.0306256

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 outlinks 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

II. OVERVIEW OF THE VARIOUS PERSONALIZATION CATEGORIES

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 user information. 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 datamining 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

Web usage data mining personalization:

The customer preference and the product association are automatically learned from click stream.

III. THREE KINDS OF INFORMATION RETRIEVAL SYSTEM

According to the architecture, there have been three kinds of information retrieval systems, which are introduced as follows.

□ Centralized search system. This system has its own data collecting mechanism, and all the data are stored and indexed in a conventional database system. Although many web search engines download web pages and provide service by thousands of servers, they all belong to this kind according to their basic architecture.

□ Metadata harvest search system. Normally Metadata is much smaller than data itself. So when we can't store all the data in a database system or need to integrate different kinds of information resource like video, PDF, web pages in a system, we can harvest the metadata from different sub databases and build a union metadata database. This database can provide the search service just like the conventional database system. Some library database systems based on OAI [10] just adopt this method.

□ Distributed search system. When the data source is large enough that even the metadata can't be efficiently managed in a database system, we can choose distributed system. Distributed information retrieval system has no its own actual record database. It just indexes the interface of sub database system. When receiving a query from a user, main system will instantly obtain the records from sub databases through their search interfaces. The limitation of this system is that the number of sub databases can't be many, otherwise the search speed can't be ensured. A famous system is InfoBus system in Stanford digital library project [11].

There are two main factors to determine the architecture of an information retrieval system, the size and diversity of data source. Normally, with the increase of size and diversity of data source, we can select centralized system, metadata harvest system and distributed search system respectively.

IV USAGE PATTERN DISCOVERING

Several research studies have been made to model individual and group behaviours and to evaluate usage patterns of different services. These models have used different sources of input data for modelling. These input data includes access log files , click trace , questionnaires , interviews and other relevant documents . In Web access log files and clicking patterns of users to the Website have been used to evaluate the usage patterns of contents of the visited Websites and to cluster the users based on their preferences for the pages from the Websites. These studies have been used to improve the Website contents and to eliminate those contents which are not being used. A similar research study has been made to predict Website user's genders, age and their ethnicity.

V WEBSITE CLASSIFICATION SCHEME

Web page classification is the process of assigning a Web page to one or more predefined category labels. In Website classification, categorization can be done based on Website's content or structure. Most of the general purpose search engines and portals use the Website classification scheme of Open Directory Project (ODP) [11]. These search engines and portals include Google, Netscafe Search, AOL Search, Lycos, DirectHit, etc. ODP is a multilingual open content directory of WWW links and is constructed and maintained by a community of volunteer editors. ODP defines 16 top level categories, which are 1: Arts, 2: Business, 3: Computers, 4: Games, 5: Health, 6: Home, 7: Kids and Teens, 8: News, 9: Recreation, 10: Reference, 11: Regional, 12: Science, 13: Shopping, 14: Society, 15: Sports and 16: World.

Since, ODP categories to which Websites visited by users , are related to activities of user behavior environments. We need to have the concepts in the classification scheme which explicitly are related to the activities of user

We can design a two step process to identify the website visited by the user

Step 1:
To identify the category of visited Websites and average time spent by the user
• For each URL visited, get the top level domain name

• First the name of Websites will be extracted by trimming the resource accessed. For example, the URL http://www.google.com/resources/xyz.abc.html will become http://www.google.com.

• Extract the top level domain name from the URL extracted. For example, the formatted URL from the above step http://www.google.com will be become google.com

Note: For all sub-domains, this process produces the same top level domain. For example, for different Websites visited including http://www.google.com, https://mail.google.com/mail/, https://plus.google.com/, google.com will be the resultant extracted top level domain
.
• After retrieving the top level domain name, this domain name will be searched in 'Content Database' for finding the different categories that the extracted top level domain has been listed.

For example, the search results obtained for Websites google.com and facebook.com are 2417 category matches for google.com and 249 category matches for facebook.com.

Step 2:

The input for this is Proxy server access log files.
• In this step, for each unique top level domain names extracted from step 1, we have mapped the categories to our proposed Website's classification scheme.

For example, for twitter.com, one category we got from ODP is Computers: Internet: On the Web: Online Communities: Social Networking: Twitter. From our mapping it can be categorized under Society → Social Network → Social Communication category.

Session:

This table contains the analysis data which includes session details. Session time is the amount of time which a user spent on Internet continuously. If the amount of time between two consecutive hits for a user is greater than the pre-defined (here 15 minutes) session time then a new session is created. This table includes the fields: Session_Id, User_ Id, Day_ History_ Id: reference to Day-history-table, Start_ Time, End_ Time and Session_ Duration.
Day-History:

This table contains the records of day history for each user. The fields included for this table are: Day_History_Id, User_Id, No_of_Sessions, Average_Session_Duration, Minimum_Session_Duration, Maximum_Session_Duration. Further, this module also populates and handles data for Website, Category, Session-Website, and Website-Category tables.

IV. PROPOSED WORK

Web users typically submit very short queries to search engines, the very small term overlap between queries cannot accurately estimate their relatedness. Given this problem, the technique to find semantically related queries (though probably dissimilar in their terms) is becoming an increasingly important research topic that attracts considerable attention.

After the survey and research, it has been found that the need of having a search engine procedure or any searching technique which gives more refined and accurate search results in any of the user defined context. As the various search engines currently present in the market may or may not give the relevant or related search results. So to fill the gap between the output of a search engine from related search results to more related and relevant search results, a technique is required.

The architecture of my proposed research work is represented by a diagram. The implementation has five modules

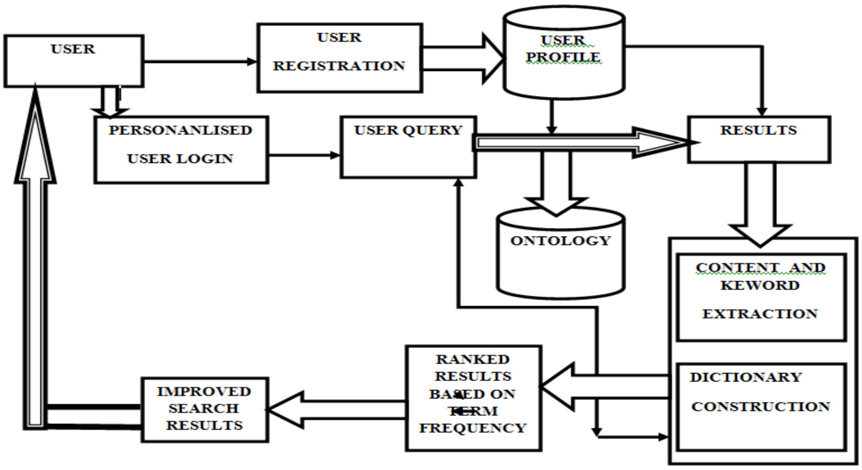
1.User Profile and Ontology Construction

2. Query mapping and search results

3. Content and keyword extraction

4. Ranking

5.Improved Search Results.



A framework for contextual information access using ontologies and demonstrated that the semantic knowledge embedded in an ontology combined with user profiles can be used to effectively tailor search results based on users' interests and preferences

V. 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. Ecommerce 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] Subha Palaneeswari M., Ganesh M., Karthikeyan T., Manjula Devi A.J., Mythili S.V., "Hepcidin-minireview", Journal of Clinical and Diagnostic Research, ISSN : 0973 - 709X, 7(8) (2013) pp.1767-1771.

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

[4] Laljee R.P., Muddaiah S., Salagundi B., Cariappa P.M., Indra A.S., Sanjay V., Ramanathan A., "Interferon stimulated gene - ISG15 is a potential diagnostic biomarker in oral squamous cell carcinomas", Asian Pacific Journal of Cancer Prevention, ISSN : 1513-7368, 14(2) (2013) pp.1147-1150.

[5] 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.

[6] Kumar S., Das M.P., Jeyanthi Rebecca L., Sharmila S., "Isolation and identification of LDPE degrading fungi from municipal solid waste", Journal of Chemical and Pharmaceutical Research, ISSN : 0975 – 7384 5(3) (2013) pp.78-81.

[7] R. Kosala and H. Blockeel, "Web Mining Research: A Survey," *SSIGKDD Explorations,ACM SIGKDD*, July 2000.
[8] Khanaa V., Thooyamani K.P., Saravanan T., "Simulation of an all optical full adder using optical switch", Indian Journal of Science and Technology, ISSN : 0974-6846, 6(S6)(2013) pp.4733-4736.
[9] A. J. Ratnakumar, "An Implementation of Web Personalization Using Web Mining Techniques," J*ournal of Theoretical and applied information technology*, 2005.
[10] Langeswaran K., Revathy R., Kumar S.G., Vijayaprakash S., Balasubramanian M.P., "Kaempferol ameliorates aflatoxin B1 (AFB1) induced hepatocellular carcinoma through modifying metabolizing enzymes, membrane bound ATPases and mitochondrial TCA cycle enzymes", Asian Pacific Journal of Tropical Biomedicine, ISSN : 2221-1691, 2(S3)(2012) pp.S1653-S1659.
[11] 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
[14] Sangeetha Rajagurusamy, Analysis of Work study in An Automobile Company ,International Journal of Innovative Research in Science, Engineering and Technology, ISSN: 2319-8753 , pp 5622-5631,Vol. 2, Issue 10, October 2013.
[15] V.G.Vijaya,Analysis of Rigid Flange Couplings ,International Journal of Innovative Research in Science, Engineering and Technology ,ISSN: 2319-8753 , pp 7118-7126,Vol. 2, Issue 12, December 2013.
[16] V.G.Vijaya ,DESIGN OF HUMAN ASSIST SYSTEM FOR COMMUNICATION ,International Journal of P2P Network Trends and Technology(IJPTT),ISSN: 2319-8753 ,pp 3687-3693,Vol. 2, Issue 8, August 2013.
[17] V.G.Vijaya, V.Prabhakaran ,Design of Human Assist System for Communication ,International Journal of Innovative Research in Science, Engineering and Technology ,ISSN: 2249-2651,pp 30-35, Volume1 Issue3 Number1–Nov2011.
[18] V.Krishnasamy, R.Kalpana devi ,Isomorphous Salts with Abnormal Water Of Hydration ,International Journal of Innovative Research in Science, Engineering and Technology,ISSN: 2319-8753 , pp 3500-3509,Vol. 2, Issue 8, August 2013.
[19] Veera Amudhan R ,Tracking People in Indoor Environments ,International Journal of Innovative Research in Science, Engineering and Technology,ISSN: 2319-8753 , pp 15996-16003,Vol. 3, Issue 9, September 2014.

DOI: 10.15680/ijircce.2015.0306256