Analysis : Web Personalization association Via Web Mining Technique

Satya Prakash Awasthi

Department of Information Technology,
Shri Venkateshwara University,
Gajraula, UP, INDIA,

Sandeep Gupta

Department of Computer Science, JIMS Engineering Management Technical Campus,
Greater Noida, UP, INDIA

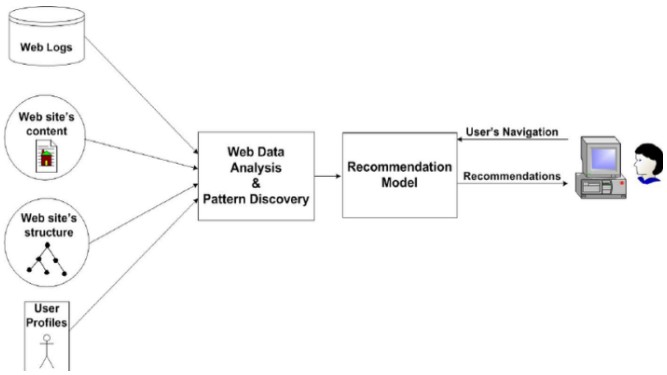
Abstract —The increase in information resources on the World Wide Web permits users to search out the data they have and navigate through multiple sites on the Net, because the WWW is a vast repository that is growing exponentially. Users typically unable to succeed in the lookout page when surfing the net.. Web personalization is proposing approach to ease the individuals from burden of data overload on net and supply them relevant information as per their needs. Web personalization is a strategy, a marketing tool, and an art. Personalization needs implicitly or explicitly collecting visitor data and leverages that data in our content delivery framework to control what information we present to our users and the way we present it.

Keywords: Web Personalization, Web Mining, Preprocessing

I. INTRODUCTION

Web Personalization is proposing approach to ease the people from burden of data overload on web t and supply them relevant information as per their requirements. The goal of web personalization using web mining is to identify interesting patterns from web usage data and recommend objects to the user that consists of products, text, and links then on. World Wide Web, the largest data base, is growing disorganized method. The web pages are linked each other, but are not logically organized. Same time probably millions of web pages are added to www and also undergo changes daily [1]. This leads information overloading. Therefore during this situation, finding required data or particular detail is tedious. Web mining consists of several techniques like, Web content Mining, Web structure mining, and Web usage mining. Which helps in filtering valuable learning from web information?

Fig. 1. Web Personalization Process[2]



THE OVERALL PROCESS OF WEB [3-8] PERSONALIZATION IN GENERAL, CONSISTS OF FOLLOWING TASKS.

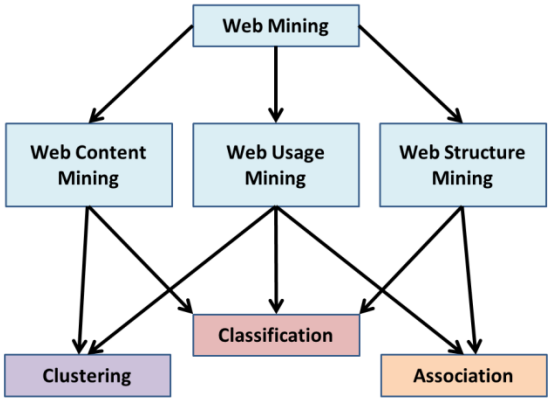
1. Data collection: During this task data is collected which might be usage information, content information, user profile information and structure information.
2. Data pre-processing: Data is then pre-processed which is necessary task to find out interesting usage patterns.
3. User profiling: It is the method of gathering user specific information, either implicitly or explicitly. The user profile will embody his/her information, interest and navigational behaviour while surfing on net.
4. Extracting useful knowledge and interesting usage patterns: The data collected is then analysed to extract useful knowledge and interesting usage patterns. There are various approaches to analyse the data which includes, content-based filtering, collaborative filtering, rule based filtering, web usage mining technique etc.
5. Pattern analysis: In pattern analysis, uninteresting rules or patterns are filtered out.
6. Personalization: It includes the actions to be carried out, recommended by such personalization systems.

II. RELATED WORK

A. *Web Mining*

The WWW is huge and growing exponentially. It contains vast amount of information that is growing and updating rapidly. Various companies, institutes, government agencies and service centers update their information regularly. The web pages do not have any standard structure and carry complex style[9-15]. Also, the web pages are organized in complex fashion than any other traditional text documents. The WWW provides its services to the varieties of web users. Net users could have completely different interests, requirements and backgrounds. Once a user searches for the information on internet, actually he/she is interested only in little portion of information.The challenges given above encourage in finding out some means to use web resources effectively, which also leads to the web mining. Most of the researchers call web mining to all strategies that apply data mining to web data. Web Mining can be defined as application of data mining techniques to extract knowledge from the web data. Mainly there are three categories to carry out web mining task: web usage mining, web structure mining and web content mining.

Fig. 2. Process of Web Mining



B. *Web Mining Techniques*

As shown in Figure 2, web mining is broadly divided into three categories according to the kinds of data to be mined: web content mining, web structure mining and web usage mining[16-22].

1. **Web content mining**: It is the task of extracting useful information from the content of web documents. The contents of web documents can be text information, some video, any image and graphs. Actually many times, those contents of web documents are in unstructured or semi-structured format and hence extracting the useful information or knowledge becomes tough and complicated. To mine the contents of web pages multimedia data mining and text mining are useful.

2. **Web structure mining**: It depends on the structure of web documents. It includes XML or hyper text markup language links/tags used in web pages. Normally various pages are linked together via HTML hyperlinks. So by studying these hyperlink connections, some useful information such as importance of the particular web page, can be found out. If the web page is linked to many other web pages, then it can be considered as an important page and can be placed in higher rank category. [WSYL09, SCDT00, Zho06, SKVP13, WYZ06, Min]. Social network analysis is the famous research done in the area of web structure mining.

3. **Web usage mining:** It is the application of data mining techniques which aims to discover interesting and frequent access patterns from web log data.

The term Web Usage Mining (WUM) was introduced by Cooley in 1997. Web usage mining (also known as web log mining) is the application of data mining techniques which aims to discover interesting and frequent access patterns from web log data [CMS97, Rat5]. The extracted interesting usage patterns and knowledge can be used in varieties of applications like system improvement, website modification/restructuring, use of caching & pre-fetching for improvement of user navigation and personalized web.

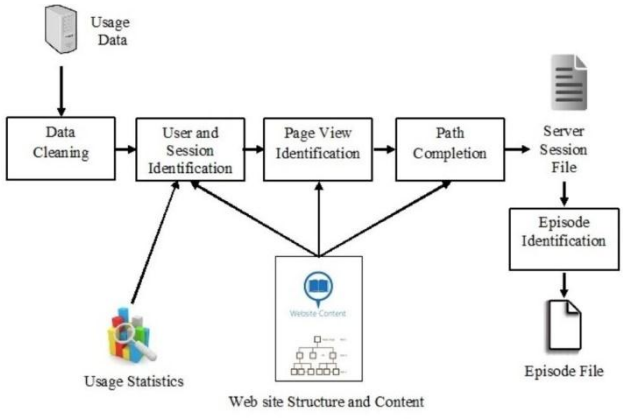
C. *Preprocessing*

The preprocessing of web logs is usually complex and time consuming. It consists of four steps:

1. Data cleaning,
2. Identification and the reconstruction of user's sessions,
3. Retrieving of information about page content and structure
4. Data formatting.

Data cleaning step consists of removing all entries/data in web logs that are irrelevant and not useful in mining process e.g. graphical page content (e.g. jpg and gif images), or the requests of robots and web spiders are considered as irrelevant and useless . Robots and web spiders related irrelevant entries are found out by referring to the user agent, or by checking the text file; robots.txt. A heuristic based approach can be used in the cases where robots send false information such as false user agent in HTTP request. In such approach, the user's sessions and robots sessions are separated.

Fig. 3. Process of Web Usage data



Sessionizationis the process of segmenting the user activity into sessions. Episode identification can be performed as a final step in pre-processing of the click stream data in order to focus on the relevant subsets of page-views in each user session. An episode is a subset or subsequence of a session comprised of semantically or functionally related page views .

Session identification step involves the identification of different users' sessions. Those sessions are identified using incomplete information form web logs. The use of proxy servers create the caching problem, which affects on session identification. So sessions can be reconstructed by using navigation oriented heuristics, time oriented heuristics or using cookies.

The last step of preprocessing is to format the data properly and then provide the formatted data for mining purpose. Data can be formatted in various ways such as; to use relational database to store data extracted from web logs, to use signature tree for indexing the logs or to use WAP-tree to store access sequence. Even a cube-like structure can be used to store session information [GS05, CMS99].

D. Pattern Discovery

Pattern discovery aims to detect interesting patterns from the preprocessed web usage data i.e. mining the data. . It includes methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition. Generally, there are many data mining techniques particularly for web personalization based on classification, clustering, sequential pattern mining, association rule discovery and statistical approaches. Among them, sequential pattern mining method is an extensively used data analysis technique in web usage mining.

There are various data mining techniques that are used in web usage mining e.g. association rule mining is used in many web usage mining applications. The aim of association rule mining is to identify association or correlation between different data items or different set of data items.

Clustering partitions a set of objects in groups (clusters), such that objects within the same group bear a closer similarity to each other, than objects in different groups A cluster is a collection of data objects where all data objects are similar with each other in same cluster but are dissimilar to the objects in other clusters . The choice of clustering algorithm depends on the application and type of data available

E. Pattern Analysis

The aim of pattern analysis is to convert discovered rules or patterns into knowledge. Here the knowledge means conceptual idea which describes the information to understanding [Sci]. It is highly dependent on a person performing the analysis. Also the exact method of analysis depends on the application for which web mining is done. For example it is done by using knowledge query mechanism, like SQL. Another method is to perform OLAP (OnLine Analytical Processing) operations using usage data. Visualization techniques, such as graphing patterns or assigning colors to different values, can often highlight the overall patterns or trends in the data. Content and structure information can be used to filter out patterns containing pages of a certain usage type, content type, or pages that match a certain hyperlink structure [23-37].

Limitations and challenges in web mining:

1. As web is extremely huge and rapidly increasing it becomes a challenging task to mine the web.
2. It becomes hard to handle unorganized ,non-standard, heterogeneous and irregular data patterns.
3. Organizing hardware and software for such a complex and extremely large processing is also challenging.
4. The information source available is rapidly changing source which gives challenge for web mining.
5. The web users belong to the various backgrounds, with completely different purpose and with different interest. Web mining needs to handle this diversity.

III. CONCLUSION

In this paper various Web Mining categories like Web content mining, web structure mining, web usage mining have been discussed. Then paper discusses about various limitations and challenges of web mining.

ACKNOWLEDGEMNT

I offer my genuine thanks to my guide Dr. Sandeep Gupta, Associate Professor (CSE Department) for his steady help, worth full direction and support amid the work. I might want to gratitude to SVU, Uttar Pradesh for giving me such stage for taking my exploration work to certain statures.

REFRENCES

[1] D. Zhou, Xuan Wu, Wenyu Zhao, Séamus Lawless, Jianxun Liu, "Query Expansion with Enriched User Profiles for Personalized Search Utilizing Folksonomy Data" IEEE Transactions on Knowledge and Data Engineering Volume: 29, Issue: 7, Pages: 1536 - 1548, 2017.

[2] C. Chen, Xiangwu Meng, Zhenghua Xu,Thomas Lukasiewicz, "Location-Aware Personalized News Recommendation With Deep Semantic Analysis", IEEE Access, Volume: 5 Pages: 1624 - 1638, 2017.

[3] S. Liang, Fei Cai, Zhaochun Ren, Maarten de Rijke, "Efficient Structured Learning for Personalized Diversification", IEEE Transactions on Knowledge

and Data Engineering Volume: 28, Issue: 11 Pages:2958 - 2973, 2016.

[4] L. Wang, Bin Wu, Juan Yang, Shuang Peng, "Personalized recommendation for new questions in community question answering", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) Pages: 901 - 908, 2016

[5] T M Veeragangadhara swamy, G T Raju, "A Novel Prefetching Technique through Frequent Sequential Patterns from Web Usage Data", COMPUSOFT An international journal of advanced computer technology, vol. 4, no. 6, June 2015

[6] M. Dhandi, Rajesh Kumar Chakrawarti "A comprehensive study of web usage mining", Symposium on Colossal Data Analysis and Networking (CDAN) Pages: 1 - 5, 2016

[7] C. Ramesh, K. V. Chalapati Rao, A. Govardhan "Ontology-based web usage mining model", International Conference on Inventive Communication and Computational Technologies (ICICCT) Pages: 356 - 362, 2017.

[8] X. Wu, Dong Zhou, Yu Xu, Séamus Lawless, "Personalized query expansion utilizing multi-relational social data" 12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP) Pages: 65 - 70, 2017.

[9] J. Hoxha, Peter Mika, Roi Blanco, "Learning Relevance of Web resources across Domains to make recommendations", 12th international conference on Machine Learning and Applications, vol. 2, pp. 325-330, 2013.

[10] F. Akhlaghian, B. Arzanian and P. Moradi"A Personalized Search Engine Using Ontology-Based Fuzzy Concept Networks", International Conference on Data Storage and Data Engineering (DSDE), Pages. 137 -141, 2010..

[11] H. Yilmaz, P. Senkul, "Using ontology and sequence information for extracting behavior patterns from web navigation logs", Data Mining Workshops (ICDMW) 2010 IEEE International Conference on, pp. 549-556, Dec. 2010

[12] K.W.-T. Leung, D.L. Lee and Wang-Chien Lee, "Personalized Web search with location preferences", IEEE 26th International Conference on Data Engineering (ICDE), Pages. 701 - 712, 2010

[13] H.-joon Kim, Sungjick Lee, Byungjeong Lee and Sooyong Kang, "Building Concept Network-Based User Profile for Personalized Web Search", 9th International Conference on Computer and Information Science (ICIS), Pages. 567 - 572, 2010

[14] J. Yu and Fangfang Liu, "Mining user context based on interactive computing for personalized Web search", 2nd International Conference on Computer Engineering and Technology, Vol.-2, Pages. 209-214, 2010.

[15] J. Lai and B. Soh, "Personalized Web search results with profile comparisons," 3rd International Conf. on Information Technology and Applications-2005, Vol. -1, Pages. 573 - 576, 2005

[16] H. Liu and V. Keselj, "Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users future requests", Data and Knowledge Engineering, Elsevier, 2007.

[17] K.-Joong Kim and Sung-Bae Cho, "A personalized Web search engine using fuzzy concept network with link structure", Joint

[18] 9th IFSA World Congress and 20th NAFIPS International Conference, Vol. 1, Pages. 81 -86, 2005.

[19] M. Kutub, R. Prachetaa and M. Bedekar, "User Web Search Behaviour," 3rd International Conference on Emerging Trends in Engineering and Technology, Pages. 549 - 554, 2010

[20] B. Ganter and R. Wille. Formal Concept Analysis: Mathematical Foundations. Springer-Verlag, 1997.

[21] M. Gorden, "Probabilistic and genetic algorithm for document retreival", communication of the ACM 31(10) (1988) 1208-1218

[22] D. Vrajitoru, "Crossover improvement for the genetic algorithm in information retreival", information processing and management 34(4)(1998) 405-415

[23] W. Fan. M. Gordan, P.Pathak, "Personalization of search engine services for effective retreival and knowledge management ", in: Proc, 2000 International conference in information system, Milwaukee, USA, August 1999.

[24] Hong Guo, Lindsay B. jack and Ashok K Nandi, " Feature Generation using Genetic programming with Application to fault classification' IEEE transaction on system and cybernetics. Vol 35, No.1(2005)

[25] Koza JR Genetic Programming as a means for Programming Computers by natural selection. Stat Comput4(2): 87-112

[26] Dr R.R Raja Laxmi, A. Sylvia Rani, " Unsupervised feature selection using Binary Bat Algorithm", 2nd International Conference of electronics and communication systems(ICECS), IEEE sponsored, 2015.451-456

[27] A.M. Robertson, P. Willet, " Generation of equifrequent groups of words using a genetic algorithm, journal of documentation "50(3) (1994) 213-232.

[28] M. Gordon, User-based document clustering by redescribing Subject description with a genetic

algorithm, Journal of the america society for information science 42(5)(1991)311-322.

[29] Anirban Mukhopadhyay, Senior member IEEE, Ujjwal Maulik, Senior Member , IEEE, Sanghamitra Bandyopadhyay, Senior Member, IEEE and Carlos Artemiocoello Coello, Fellow, IEEE "A Survey of multiobjective Evolutionary Algorithm for Data Mining: Part-II" IEEE Transaction of Evolutionary Computation, Vol. 18, 2014.

[30] G. Piatetsky – Shapiro, U.Fayyad and P.Smith from Data Mining to Knowledge Discovery "An Overview Advance in knowledge Discovery and Data Mining" Pages 1-35, AAAI/MIJ/1996

[31] Razavian A S, Azizpour H, Sullivan J, et al. 2014,"CNN features Off the Shelf: An Astounding Baseline for Recognition. In Proceeding of the 2014 IEEE Conference on computer Vision and Pattern Recognition workshops "

[32] http://www.image-net.org/challenges/LSVRC/2012/supervised

[33] Yang XS. (2010) A New Metaheuristic Bat-Inspired Algorithm. In: González J.R., Pelta D.A., Cruz C., Terrazas G., Krasnogor N. (eds) Nature Inspired Cooperative Strategies for Optimization (NICSO 2010). Studies in Computational Intelligence, vol 284. Springer, Berlin, Heidelberg

[34] P. Vora and B. Oza, "A survey on k-mean clustering and particle swarm optimization," Int. J. Sci.Modern Eng, vol. 1, pp. 24-26, 2013

[35] S. Cho, J. Lee, A human-oriented image retrieval system using interactive genetic algorithm,IEEE Transactions on System, Man and Cybernetics. Part A: Systems and Humans 32 (3)(2002) 452–458.

[36] S. Kato, S. Iisaku, An image retrieval method based on a genetic algorithm, in: Proc. Twelfth International Conference on Information Networking (ICOIN-12), 1998, pp. 333–336.

[37] Z. Stejic, Y. Takama, K. Hirota, Genetic algorithm-based relevance feedback for image retrieval using local similarity patterns, Information Processing & Management 39 (2003)1–23.