International Journal of Information Processing, 8(1), 40-52, 2014
ISSN : 0973-8215
IK International Publishing House Pvt. Ltd., New Delhi, India

A Framework for Preprocessing Web Log in the Data Warehouse Environment for Web User Behavior Analytics

B N ShankarGowda[a], Vibha Lakshmikantha[b], K R Venugopal[c], L M Patnaik[d]

[a]Department of Computer Science and Engineering, Bangalore Institute of Technology, Bangalore 560 004 India, Contact: bnsgowda@gmail.com

[b]Department of Computer Science and Engineering, BNMIT, Bangalore 560 070 India,

[c]Principal, University Visvesvaraya College of Engineering, Bangalore University, Bangalore 560 001 India,

[d]Indian Institute of Science, Bangalore, India.

The rapid growth of World Wide Web (www) in the recent times is redefining the economies of nations. The www is the largest available source of information and the users interaction with the www is increasing exponentially. The user interests reflect the user background and topics of interests which the users leave behind the navigation traces of their interaction, which is extracted and forms basis for the analytics of user behavior. A better understanding of the users ever evolving behavior, habits, needs and interests will allow the organization to benefit in a big way. The objective of web data analytics is to process and examine large amounts of data of a variety of types (big data) to uncover hidden patterns; unknown correlations and other useful information so as to aid organisations decision making process. Web Usage Mining deals with discovering user access patterns from web log data by extracting, modeling and analyzing the behavioral patterns of user interaction with the web. This collection of large and complex data set is difficult to process using the conventional DBMS tools or traditional data processing applications or desktop statistics and visualization packages, as they require massively parallel software running on tens, hundreds of servers. Even though, several Data mining techniques are used to uncover the hidden information in the web, there is scope for improving the techniques for the effective analysis of ever evolving user behavior. This paper proposes an efficient methodology for analyzing user behavior by building a DWH and integrating it with the Data mining framework. The primary aspect of web data analytics is, to obtain a good dataset in which data is clean, accurate, predictive, timely, accessible and complete. A good dataset is obtained by preprocessing the Web Log in Data warehouse environment and also enhances the performance, throughput, scalability and multi-dimensional analysis economically.

Keywords : EWIC: Enterprise Wide Information System, MDD: Multi Dimensonal Data, ODS: Operational Data Source, SOM: Self-Organizing Maps, WUM: Web Usage Mining.

1. INTRODUCTION

The World Wide Web is the largest available source of information and the users interaction with the web is increasing exponentially which has generated large amount of complex data of a variety of types related to the users interactions with the Web sites. This data is recorded in the servers log files or web log data and usually referred as Web Usage Data (WUD). The users leave behind the navigation traces of their interaction, which is extracted and is a basis for the user behavior analytics to uncover hidden patterns, their correlations and other useful information. The users interests can provide substantial clues for document efficient Information Retrieval (IR) that reflects generally the user background and topics of their interests. The fast pace of data accumulation and the big data has made it imperative for automated

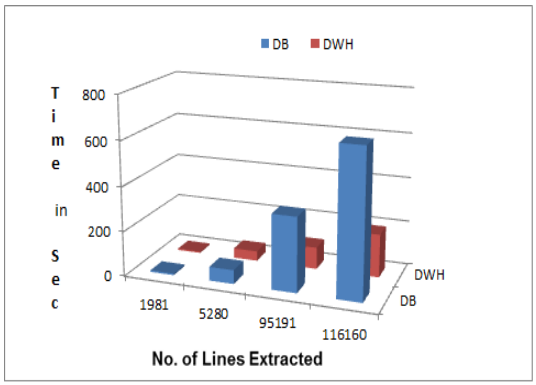

Figure 8. Extracting 4 variables using DB and DWH

Table 1
Extracting all the variables from log file

No. of Lines	Scripts	Database	DWH
1981	121	7	3
5280	401	20	7
95191	12668	364	62
116160	15568	431	98
232320	33498	851	223

Table 2
Extracting 4 variables using Scripts, DB and DWH

No. of Lines	Scripts	Database	DWH
1981	21	4	3
5280	64	7	5
95191	3656	62	44
116160	4645	331	98
232320	10616	665	193

8.1. Advantages of DWH Framework

The advantages of DWH approach against the operational databases using a query language like SQL are a plenty. The DWH and data mining framework provides with several optimizations techniques for computing multiple aggregates that are not supported in case of TPS or operational databases and we need to write complex procedures to perform the aggregate operation on the data for analysis. The advantages of using the proposed methodology are summarized as shown in Table 3.

Table 3
Advantages of DWH and DM framework

Parameter	Scripts/DB	DWH
MDD	X	Easy
CUBE()	X	Easy
RollUp()	Tedious	Easy
Drilldown()	Tedious	Easy
Rank()	Tedious	Easy
Dense Rank()	Tedious	Easy
Slice()/Dice()	X	Easy
Pivot()	X	Easy
Aggregation()	Tedious	Easy
Summarization()	Tedious	Easy

9. CONCLUSIONS

The experimental results show that building DWH and integrating it with the Data mining tools yields better performance when compared to traditional approaches where the data set is prepared depending on the goals of the analysis, the data set needs to be transformed and aggregated at different levels of abstraction. The experimental results of data sourcing Acquisition, cleanup, transformation process confirms the reduction in the size of Server Log file considerably and also arrives at a good dataset that aids in enhancing the performance of data analysis.

REFERENCES

1. Xiaohui Yan, Jiafeng Guo and Xueqi Cheng. Context-Aware Query Recommendation by Learning High-Order Relation in Query Logs, *In the 20th ACM international conference on Information and knowledge management, Glasgow, Scotland, UK*, pages 2073–2076, October 2428, 2011.
2. E Frias-Martinez and V Karamcheti. A Customizable Behavior Model for Temporal Prediction of Web User Sequences, *In the Fourth International Workshop WEBKDD, Edmonton, Canada*, pages 66–85, July 23, 2002.
3. M Spiliopoulou, B Mobasher, B Berendt and M Nakagawa. A Framework for the Evaluation of Session Reconstruction Heuristics in Web-Usage Analysis, *In INFORMS Journal on Computing*, 15(2):171–190, 2003.

4. Cooley R, Mobasher B and Srivastava J. Web Mining: Information and Pattern Discovery on the WWW, *In Ninth IEEE International Conference on Tools with Artificial Intelligence (ICTAI 97), Newport Beach, CA*, pages 558–567, 03-08 Nov, 1997.
5. Cooley R, Tan P N and Srivastava J. Discovery of Interesting Usage Patterns from Web Data, *In the International WEBKDD99 Workshop, Springer Berlin Heidelberg, San Diego, CA, USA*, pages 163–182, August 15, 1999.
6. Natheer Khasawneh and Hien Chung Chan. Active User-Based and Ontology-Based Web Log Data Preprocessing for Web Usage Mining, *In Proceedings of IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies, Hongkong*, pages 325–328, 2006.
7. Zhang Kejun, Zhao Geng, Li Boqun and Chen Zuo. Research of Distributed Web Interest Conversion Pattern Mining, *In the IEEE International Conference on Control and Automation, Guangzhou, CHINA*, pages 1828–1832, May 30 - 01 June, 2007.
8. Qiang Yang, Joshua Zhexue Huang and Michael Ng. A Data Cube Model for Prediction-Based Web Prefetching, *In the Journal of Intelligent Information Systems*, 20:11–30, 2003.
9. Hussain T, Asghar S and Masood N. Web Usage Mining: A Survey on Preprocessing of Web Log File, *In the International Conference on Information and Emerging Technologies (ICIET), Karachi, Pakistan*, pages 1–6, 2010.
10. Vijaya Kumar T and Guruprasad H S. Clustering Web Usage Data Using Concept Hierarchy and Self Organizing Map, *In International Journal of Computer Application*, 56:38–44, 2012.
11. Kewem Liu. Analysis of Preprocessing Methods for Web Usage Data, *In the International Conference on Measurement, Information and Control (MIC), Harbin, China*, pages 383–386, 18-20 May,2012.
12. Milija Suknovic, Milutin Cupic and Milan Martic. Data Warehousing and Data Mining - A Case Study, *In the Yugoslav Journal of Operations Research*, 15: 125–145, 2005.
13. Ling Zheng, Shuo Cui, Dong Yue and Xinyu Zhao. User Interest Modeling Based on Browsing Behavior, *In the Third International Conference on Advanced Computer Theory and Engineering (ICACTE), Chengdu, China*, pages 455–458, 2010.
14. Tanasa D and Trousse B. Data Preprocessing for WUM, *In the Potentials, IEEE*, pages 22–25, Sept-2004.
15. Olivia Rud C. Data Warehousing for Data Mining: A Case Study, *In the SAS Users Group International Conference, SUGI-25, Indiana, USA*, pages 119–125, 2000.
16. Huang Hao, Jiang Dan and Huang Jianqing. Separating Interleaved User Sessions from Web Log, *In the International Conference on Network Computing and Information Security (NCIS), Guilin, China*, pages 152–156, 2011.
17. Raju G T, Yogish and Manjunath T N. The Descriptive Study of Knowledge Discovery from Web Usage Mining, *In the IJCSI International Journal of Computer Science Issues*, 8:225–230, 2011.
18. Bamshad Mobasher. The Adaptive Web, *Springer, Germany*, 2007.
19. Jiang Chang bin and Chen Li. Web Log Data Preprocessing Based on Collaborative Filtering, *In the second International Workshop on Education Technology and Computer Science, Wuhan, China*, pages 118–121, 2010.
20. Choi, Jinhyuk and Lee Geehyuk. New Techniques for Data Preprocessing Based on Usage Logs for Efficient Web User Profiling at Client Side, *In the IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies, Italy*, pages 54–57, 2009.
21. Bamshad Mobasher. chapter 12: Web Usage Mining in Data Collection and PreProcessing, *In the 13th ACM SIGKKD International Conference on Knowledge Discovery and Data Mining (KDD 2007), San Jose, CA, USA*, pages 450–483, Dec-2007.

B N Shankar Gowda is currently working as Associate Professor, Department of Computer Science and Engineering, Bangalore Institute of Technology, VTU, Bangalore. He obtained his Bachelor of Engineering from University Visvesvaraya College of Engineering, Bangalore University. He received his Masters' degree in Information Technology from

Queensland University of Technology, Brisbane, Australia.

Vibha Lakshmikantha is currently the Professor, Department of Computer Science and Engineering, BNM Institute of Technology, VTU, Bangalore. She obtained her Bachelor of Engineering Degree and her Masters degree in Electronics and Communication from University Visvesvaraya College of Engineering,Bangalore University. She was awarded Ph. D in Computer Science from Dr. MGR Research and Educational Institute, Chennai, India.

Venugopal K R is currently the Principal, University Visvesvaraya College of Engineering, Bangalore University, Bangalore. He obtained his Bachelor of Engineering from University Visvesvaraya College of Engineering. He received his Masters degree in Computer Science and Automation from Indian Institute of Science Bangalore. He was awarded Ph.D in Economics from Bangalore University and Ph.D in Computer Science from Indian Institute of Technology, Madras. He has a distinguished academic career and has degrees in Electronics, Economics, Law, Business Finance, Public Relations, Communications, Industrial Relations, Computer Science and Journalism. He has authored and edited 39 books on Computer Science and Economics, which include Petrodollar and the World Economy, C Aptitude, Mastering C, Microprocessor Programming, Mastering C++ and Digital Circuits and Systems *etc.*. During his three decades of service at UVCE he has over 400 research papers to his credit. His research interests include Computer Networks, Wireless Sensor Networks, Parallel and Distributed Systems, Digital Signal Processing and Data Mining.

L M Patnaik is currently Honorary Professor, Indian Institute of Science, Bangalore, India. He was a Vice Chancellor, Defense Institute of Advanced Technology, Pune, India and was a Professor since 1986 with the Department of Computer Science and Automation, Indian Institute of Science, Bangalore. During the past 35 years of his service at the Institute he has over 700 research publications in refereed International Journals and refereed International Conference Proceedings. He is a Fellow of all the four leading Science and Engineering Academies in India; Fellow of the IEEE and the Academy of Science for the Developing World. He has received twenty national and international awards; notable among them is the IEEE Technical Achievement Award for his significant contributions to High Performance Computing and Soft Computing. His areas of research interest have been Parallel and Distributed Computing, Mobile Computing, CAD for VLSI circuits, Soft Computing and Computational Neuroscience.