A Comprehensive Study of Web Usage Mining Tools: Webminer and WebKIV

Varun Malik[1], Dr. Sanjay Singla[2], Dr. Sawtantar Khurmi[3]

[1]Phd Scholar – Punjab Technical University, [2]Professor - GGSCOMT Kharar (Punjab), [3]Professor - BMSCE Muktsar (Punjab)

Abstract: Web usage mining is a branch of data mining through which we uncover interesting data over the world wide web. Web mining is further classified into web usage mining that deals with interesting patterns and user behavior over the web. In this paper two major tools of web usage mining i.e. WEBMINER and WebKIV, have been studied, through which the authors explain how these tools work and how data is retrieved from the web server, which is then converted into interesting patterns. The authors summarize the characteristics of both WEBMINER and WebKIV tools and present a comparison of these tools w.r.t. web mining.

***Keywords** - Web Mining; Web usage Mining; WEBMINER; WebKIV;*

I. INTRODUCTION

At present time, data explosion is taking place on the web as users interacting with websites are creating and retrieving data and new users are added daily. When users surf a particular website, the data being created or retrieved is observed by the web master and this helps him to meet the requirements of users.[23]Everyday there is enormous amount of data accessed or stored in homogenous or heterogeneous database for data analytics and knowledge about the user decision making. Activity of the internet users generates a data in web server log files for information about the visitors.[3] According to [6] Web mining has become an interesting subject of concern. This term was the first termed by the Oren Etzioni in 1996.[1] Nowadays various research communities take interest in the data analytics to discover the interesting pattern from the world wide web data. Web mining can be further classified into sub categorization accordingly and discussed further in this paper.

II. WEB MINING

Web mining is a sub-division of data mining techniques focusing on the World Wide Web as the main data source. The results of data mined from the WWW may consist of a collection of facts which web pages are required to contain, and these may include unstructured text, structured content such as lists and tables, and images, video and audio. The objective for techniques of mining data is to extract information from web data including documentation, hyperlinks and web site access log files, etc. Further, web mining can be categorized into three terms.[9]

i. Web Content
ii. Web Structure
iii Web usage

Web content mining is a branch of web mining which deals with the data discovered by the search engine and extraction of useful knowledge from the hyperlinks worldwide. This information can further be used by the webmaster for making decisions which require intelligence. These tasks are usually not carried out using traditional data mining techniques. Web content mining describes the uncovering of meaningful patterns from web content in the form of data or documents.[9]Web structure mining is used for extraction of meaning information or data from the web using hyperlinks. It can be further divided into external structure mining, internal mining and URL mining. Web usage mining consists of discovering and analyzing patterns in an automated manner from related data generated as a result of interaction of users over the web, i.e., from websites and web server log data. Basically, the aim of web usage mining is to model the data, and analyze the interaction patterns of users. These patterns are visualized as structures which are regularly accessed by users having common interests.

III. WEB USAGE MINING

Web usage mining is a technique of data mining used to uncover hidden patterns from web usage data, so as to interpret and fulfill the requirements of web-based applications.[1] In web usage mining the host records the location and other general information present in browsing data of users over the web. Web usage mining can be revealed further depending on the data used. Three types of processing are included under web usage mining i.e.

i. Preprocessing
ii. Pattern discovery
iii. Pattern analysis

Figure 1 shows the process of Web usage mining.

i. Pre Processing

The data related to the actions of users is contained in log files on the server. This cannot be used for mining purposes..To make it useful it needs preprocessing before the pattern discovery phase. The preprocessing involves three steps.

The gathered data is cleaned, which implies that non-textual content is filtered out. Secondly, the sessions belonging to various users are pinpointed. A session is thus a set of related activities carried out by one user whenhe navigates through a given site. It is adifficult process to identify the sessions from the collected data, because it is not feasible that all the information needed will be found in server logs.[10]

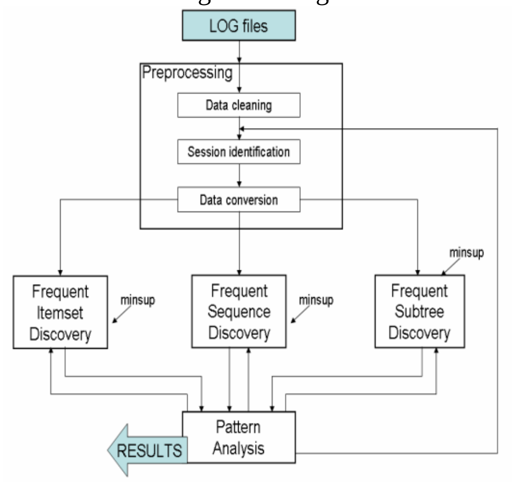

Figure1 Process of Web Usage Mining[10]

ii Pattern Discovery

This is the second step of preprocessing of data in web usage mining in which the objective of pattern discovery is to cluster the behavior of user and the pattern navigation. Clustering play a very crucial role in analyzing the data and behavior of the users in the website.. To find out the user behavior in the pre-processing, different algorithms like Weighted Fuzzy Possibilistic C-Means Algorithm can be used. [3]

iii Pattern Analysis

This is the third and last step of preprocessing and is termed as the goal of the process. In this step, patterns are elicited from the server log files and the data which is not useful is rejected.[9]

IV.WEB USAGE MINING TOOLS

i. WEBMINER

The architecture of WEBMINER system[1]is shown in Figure 1. It shows the bifurcation of web data into two main portions. The first portion covers the process of converting the web data into a suitable form. This process covers pre-processing, identifying transactions and data integration elements. The second portion consists of uncovering of association rules and patterns.

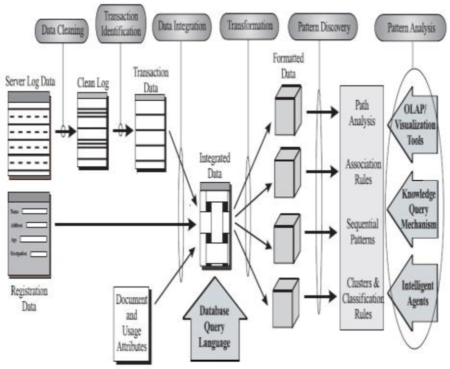

Figure2 Web Usage Mining Architecture[23]

The first step to be performed is data cleaning. Low level data integration tasks are executed at this level i.e. merging several log files, assimilating referrer logs, etc. Entries from data cleaning logs are partitioned into logical clusters using methods for identifying transactions. The target of this identification step is to design consequential groups of references for every web user. Once the transformation phase, which is dependent on the domain, is completed in Web Miner the resultant transaction data must be formatted for carrying out mining task as part of the data model. For example, the format meant for sequential pattern mining is non-identical with the format for the data for discovery of association rules [1].

ii. WEBKIV

WebKIV consists of four basic components based on its architecture.[2]

Data Collection: In WEBKIV, mainly three types of data collection are performed. In the first phase, data is collected from websites with the help of web crawlers using the breadth-first traversal algorithm. Web log data is accessed from the web server log files directly, while the mined web data is collected from web servers with the help of mining techniques. [2]

Data Parsing: To create the graph representation of aweb site a radial tree algorithm is used. Each level is allocated around a central focused node in a hierarchical acyclic tree in the radial tree algorithm.

Data Transformation. N WebKIV represents web data attributes into three shapes, namely, node representation, line representation and dot representation.

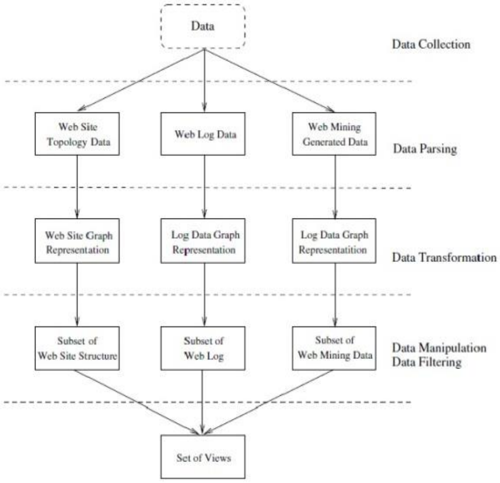
Data
Data Collection
Web Site Topology Data
Web Log Data
Web Mining Generated Data
Data Parsing
Web Site Graph Representation
Log Data Graph Representation
Log Data Graph Representatition
Data Transformation
Subset of Web Site Structure
Subset of Web Log
Subset of Web Mining Data
Data Manipulation
Data Filtering
Set of Views

Figure 3

Data Manipulation and Data Filtering: In the WebKIV system a number of data operators are applied to provide flexibility in manipulating the data over the web.

V. COMPARISON OF WEBMINER AND WebKIV TOOL

WEBMINER and WEBKIV are two tools used for analyzing web usage mining. The comparison of functionality and comparison of both tools and their characteristics are discussed here. WEBKIV is a tool used for the visualization of data in 2-D format while WEBMINER is a tool used to process or clean up the data from web server log files. Both tools are used for pattern discovery in web usage mining.

According to [6], the WEBMINER tool is a multi-purpose framework for analyzing web usage. The sub-category of data mining, namely, web mining), is an elicitation of association rules, patterns, and relationships from data collected in large depositories. The comparison shows the working functionality of WEBMINER and WEBKIV tools during different phases of data processing

VI. CONCLUSION

The WEBKIV web usage mining tool is used as a visualization tool applied on the WEBMINER architecture at the stage of pattern analysis for pattern discovery. It provides the essential functionality of visualizing in the three dimensions of either dynamic or static, large v/s small scale and individual v/s aggregate. In the WEBKIV system, to construct the

Table 1: Comparison of WEBMINER and WebKIV

PARAMETER	WEBKIV	WEB MINER
Work Layout	designed to analyze simultaneous visualization of structure, content and navigation	general architecture for mining Web data
Design	several visualization strategies are combined from tools using other web visualization techniques, a single method of visualizing web structure is available	Architecture splits Web mining activity into 2 main sections. The first section covers the process of converting the Web data into a transactional form. The second part covers the approach, which is independent of any particular domain ,to carry out data mining and, i.e. the elicitation of association rules and patterns.
Usage	Provides tools for visualizing patterns of different scales. Availability of techniques for visualizing user navigation patterns.	Data cleaning is firstly performed in the process of Web usage mining. Next step is to partition logged information into logical clusters by transaction identification modules
Identification	support primary web mining research	Identifying and dividing transactions into several small transactions or integrating small transactions into larger ones.

Process	Visualize web navigation structures and compare the patterns so constructed by applying machine learning techniques to improve navigation.	Data transformation based on the domain is completed, then resultant transaction data is formatted to finalize the data model
Pattern Discovery	Uses 2D plane to construct web site structure using a radial tree algorithm	user analyst is able to exercise greater control over the process of discovery using query mechanism by specifying different constraints.

Table 2: Comparison of WEBMINER and WEBKIV tools during data processing

Type	**WEBMINER**		**WEBKIV**	
	Feature	**FUNCTION**	**Feature**	**FUNCTION**
DATA CLEANING	Low level task is performed	Combining multiple logs, incorporating referred logs etc	Three different kind of data collected	Log data reclaimed directly from the log files
Transaction Identification	Partitioned into logical clusters	Creating meaningful clusters, divide large cluster into smaller ones	Identifying the data either topology, web log or web mining data	Applied radial tree algorithm
Data Integration	Merging small transaction into larger ones	Match input and output of transaction in any order		
Transformation	Resultant transaction data formatted	Conform the data model of appropriate data mining	**Node representation** Square is used to represent the Web page **Line representation** Access two attributes : width and color **Dot representation** Web surfer is represented by dot.	radial tree algorithm used to distribute page nodes Identify user parameters strokes indicate aggregate Dot will move during animation process
Pattern Discovery	Mapping sequential mapping	Analyst have more control	Number of data operators	Provide flexibility fc manipulate web data lik zooming, panning, subtre generation

website structure a radial tree algorithm is used in a two dimensional visualization. In the WEBMINER architecture for pattern analysis, the visualization tool WEBKIV is used for visualizing pattern discovery in a 2-D plane with the help of the radial tree algorithm. Zooming and panning techniques are applied in WEBKIV to view pattern discovery results in web mining.

As the web 2.0 spreads across the Internet, experience enhanced by using applications, and convenience extended by overcoming geographical restrictions, web usage logs are becoming a goldmine for researchers. User behavior analysis in varying areas, is usefulness for enterprises attempting to make strategic decisions. This is a common framework for working of web usage mining in which log files are extracted from the web server, a preprocessing technique is applied and datasets formed as a result of pattern analysis, pattern discovery and user statistics, are used for different applications.

VII. REFERENCES

[1]. Jaideep Srivastava, PrasannaDesikan, VipinKumar,(2005) ”Web Mining—Concepts Applications, and Research Directions”inFoundations and Advances in Data Mining”,2005

[2]. YongheNiu, Tong Zheng, Jiyang Chen, Randy Goebel,(2003) “WebKIV : Visualizing Structure and Navigation for Web Mining Applications”,at IEEE/WIC International Conference on Web Intelligence, (WI 2003), 13-17 October 2003, Halifax, Canada

[3]. Vellingiri J., S. Kaliraj, S. Satheeshkumar and T. Parthiban,(2015)"A Novel Approach for User Navigation Pattern Discovery and Analysis for Web Usage Mining",at Journal of Computer Science 2015, 11 (2): 372.382DOI: 10.3844/jcssp.2015.372.382

[4]. Dr Sanjay Kumar Dwivedi, BhupeshRawat,(2015) "A Review Paper on Data Preprocessing: A Critical Phase in Web Usage Mining Process" at 978-1-4673-7910-6/15/$31.00_c 2015 IEEE

[5]. SaloniAggarwal and VeenuMangat,(2015) "Application Areas of Web Usage Mining" at Fifth International Conference on Advanced Computing & Communication Technologies, 2015

[6]. ChhaviRana,(2012)"A Study of Web Usage Mining Research Tools"atInt. J. Advanced Networking and Applications Volume:03 Issue:06 Pages:1422-1429 (2012) ISSN: 0975-0290.

[7]. https://www.techopedia.com/definition/15634/web-mining

[8]. M. Craven, S. Slattery and K. Nigam, (1998)"First-Order Learning for Web Mining", In Proceedings of the 10th European Conference on Machine Learning, Chemnitz, 1998.

[9]. V.Anitha,Dr.P.Isakki,(2016)"A Survey on Predicting User Behavior Based on Web Server Log Files in a Web Usage Mining" at 978-1-4673-8437-7/16/$31.00 ©2016 IEEE

[10]. RenátaIváncsy, IstvánVajk,(2006) "Frequent Pattern Mining in Web Log Data"atActaPolytechnicaHungarica Vol. 3, No. 1, 2006

[11]. Subhi Jain RuchiraRawat Bina Bhandari,(2017)"A Survey Paper on Techniques and Applications ofWeb Usage Mining" at International Conference on Emerging Trends inComputing and Communication Technologies (ICETCCT),2017

[12]. Musale V.,Chaudhari D.,(2017)"Web Usage Mining Tool by Integrating SequentialPattern Mining with Graph Theory by VinayakMusale, DevendraChaudhari" at 1st International Conference on Intelligent Systems and Information Management (ICISIM), 2017

[13]. M. El Asikri, S. Krit, H.Chaib , M. Kabrane, H. Ouadani, K. Karimi,K.Bendaouad and H. Elbousty,(2017) "Mining the Web for learning ontologies: state of artand critical review"atICEMIS(978-1-5090-6778-7/17/$31.00 ©2017 IEEE) 2017, Monastir, Tunisia

[14]. Ms.Aparna M. Katekarat,(2017)"Improving the Effectiveness of Short TextUnderstanding by Using Web InformationMining"atProceedings of the IEE, International Conference on Computing Methodologies and Communication(ICCMC),2017

[15]. BrijendraSingh, Hemant Kumar Singh,(2010) "WEB DATA MINING RESEARCH: A SURVEY"at 978-1-4244-5967-4/10/$26.00 ©2010 IEEE

[16]. V.Chitraa, Dr. Antony SelvdossDavamani,(2010) "A Survey on Preprocessing Methods for Web Usage Data" at (IJCSIS) International Journal of Computer Science and Information Security,Vol. 7, No. 3, 2010

[17]. Aditya S P, Hemanth M, Lakshmikanth C K, Suneetha K R,(2017) "Effective Algorithm for Frequent Pattern Mining"at International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017),2017

[18]. M.Sathya,Dr.P.Isakki@Deviat,(2017)"AprioriAlgorithm on Web Logs for MiningFrequent Link"atIEEE International Conference On Intelligent Techniques In Control , Optimization and Signal Processing,2017

[19]. Dr.K.Shyamala, S.Kalaivani,(2017)"An Effective Web Page Reorganization throughHeap Tree and Farthest First ClusteringApproach"atIEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI-2017)

[20]. SunHao, ShenZhaoxiang, ZhangBingbing,(2017)"A User Clustering Algorithm on Web Usage Mining" atFirst International Conference on Electronics Instrumentation & Information Systems (EIIS),2017

[21]. DoruTanasa and Brigitte Trousse,(2004)"Advanced Data Preprocessing for Intersites Web UsageMining"at 1094-7167/04/$20.00 © 2004 IEEE Published by the IEEE Computer Society, March 2004

[22]. Monika Dhand, Rajesh Kumar Chakrawarti,(2016) "A Comprehensive Study of Web Usage Mining"at Symposium on Colossal Data Analysis and Networking (CDAN), 2016

[23]. R.Cooley, B.Mobasher_ and J Srivastava,(1997)" Web Mining Information and Pattern Discovery on world wide web"at Proceedings Ninth IEEE International Conference on Tools with Artificial Intelligence), 1997