# STUDY ON A FRAMEWORK TO COMPARE WEB MINING TYPES

*Vinodkumar.K*[1], *Kathiresan.V*[2]

## Abstract

In the present day the study of comparison of web mining has been made so as to assist the user about their application area. It is mainly concerned with the web content, web structure and web usage of data mining. As the data on the web is endlessly growing day by day so, web mining become more essential to represent a conclusion from the huge data available on web. In web mining non trivial pattern and useful information are retrieved from the web data. Web Mining is the application of data mining techniques to discover patterns from the Web. Web mining consists of three types namely Web usage mining, Web content mining and Web structure mining. This paper concentrate on these three types and a comparative study has been made out to find which is more effective based on various parameters and it also compare web mining with data mining.

***Keywords :*** World Wide Web, web mining, web content mining, web structure mining, web usage mining.

## I. Introduction

From the time when the web mining was introduced, a lot of improvements have been made. The simplicity and speed with which business transactions can be passed out over the Web has been an input in the swift development of electronic commerce. With the explosive growth of information sources available on the World Wide Web and the fast increasing pace of adoption to Internet business, the Internet has evolved into a gold mine that contains or dynamically generates information that is beneficial to E-Commerce business. Two different approaches were taken in primarily defining Web mining. First was a „process-centric view, which defined Web mining as a sequence of tasks [1]. Second was a data-centric view, which defined Web mining in terms of the types of Web data that was being used in the mining process [2]. Here data-centric view of the data which has been followed as detailed below:"It is the application of data mining technique to extract useful data from the web i.e. web content data, web structure data and web usage data."Web mining is achieved first by reporting visitors traffic information based on Web server log files and other source of traffic data. Web server log files were used initially by the webmasters and system administrator for the purposes of "how much traffic they are getting, how many requests fail, and what kind of errors are being generated", etc. However, Web server log files can also record and trace the visitor's on-line behaviors. For example, after some basic traffic analysis, the log files can help us answer questions such as "from what search engine are visitors coming? What pages are the most and

[1] Research Scholar Department of MCA, RVS Arts and Science College, Sulur, vinodhsml@gmail.com.

[2] Asst, Professor, Dept. of MCA, RVS Arts and Science College, Sulur, Coimbatore.

least popular? Which browsers and operating systems are most commonly used by visitors?"The attention paid to Web mining, in research, software industry, and Web-based organizations, has led to the accumulation of a lot of experiences. It is our attempt in this paper to capture them in a systematic manner, and identify directions for future research. The rest of the paper is organized as follows: Section II. introduces Web Mining. Section III. describes related work. Section IV. describes web mining taxonomies Section V. gives a comparison among the three web mining types and web mining and data mining using various parameters like view of data, method used, scope, tasks etc. Section VI. concludes the paper and Section VII. presents future scope of the work.

## II. Webmining

Web mining is one of the Data Mining techniques that automatically discovers or extracts the information from web documents. A natural combination of World Wide Web and data mining sometimes referred to as web mining. It consists of following tasks [3]:

1. Resource finding: It involves the task of retrieving intended web documents. It is the process by which we extract the data either online or offline resources available on web.

2. Information selection and pre-processing: It involves the automatic selection and pre processing of specific information from retrieved web resources. This process transforms the original retrieved data into information. The data is transformed into useful information by using suitable transformation. The transformation could be renewal of stop words, or it may be aimed for obtaining the desired representation such as finding particular format of data.

3. Generalization: It automatically discovers general patterns at individual web sites as well as across multiple sites. Data Mining techniques and machine learning are used in generalization

4. Analysis: It involves the validation and interpretation of the mined patterns. It plays an important role in pattern mining. A human plays an important role in information on knowledge discovery process on web

## III. Related Work

The focal point of the literature survey is to study or collecting information about web mining which is used to extract useful information from the web. Alexandras Nanopoulos, Dimitris Katsaros and Yannis Manolopoulos [9] focused on web pre-fetching because of its importance in reducing user perceived latency present in every web based application. From the fame of web, there is heavy traffic in the Internet and the result is that there is delay in the response. To perform any website evaluation, web visitors' information plays an important role, in order to assist this, many tools are available. Li, L,Zhang and C. And Zhang [10] expressed that Web Mining is a popular technique for analyzing website visitor s behavioral patterns in e-service systems. Jian Pei,J. Han, B.Mortazavi and Hua Zhu[11] found that Web Log Mining helps in extracting interesting and useful pattern from the Log File of the sever. The documents which are most frequently accessed by users can be placed near the home page of the website. Manoj Manuja and Deepak Garg [12] suggested the development of web mining techniques such as web metrics and measurements, web service optimization, process mining etc. will enable the power of World Wide Web to be realized. Jing Wang and others [13] found that weaknesses of both frequency and utility can be overcome by the model of General Utility Mining. Miller & Remington [14] exposed that the structure of linked pages has decisive impact factor on the usability. Geeta and others [15] suggested the number of pages at a particular level, the number of forward links and the number of backward link to a particular web page reflect the behavior of visitors to a specific page in the website. However Garofalakis [16] pointed out that the number of hit counts can be calculated from log file is not the best indicator of page popularity. Geeta & others [15] suggested that the topology of the website plays an important role in addition to log file statistics to help users to have quick response. Jia-Ching and others [17] found that Web Usage Mining helps in discovering web navigation patterns mainly to predict navigation and improve website management. Lee and others [18] proved that the web behavioral patterns can be used to improve the design of the website. These patterns also could help in improving the business intelligence.

## IV. Web Mining Taxonomy

According to analysis targets Web mining is broadly classified into three types based data to be mined:

1. Web Content Mining: Web content mining is the process of extracting useful and valuable information from the contents of web documents. Content data is the collection of data from which a web page is designed. It may consist of text, images, audio, video, or structured records such as lists and tables [7]

2. Web Structure Mining: Web structure mining is the process of discovering structured information from the web. The structure of a typical web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. This can be further divided into two kinds based on the kind of structure information used. First is Hyperlink Structure that connects a location in a web page to a different location, either within the same web page or on a different web page. A hyperlink that connects to a different part of the same page is called an intra-document hyperlink, and the hyperlink that connects two different web pages is called an inter-document hyperlink. And the second one is Document Structure that contains the content within a Web page that can also be organized in a tree structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents [5].

3. Web Usage Mining: Web usage mining is the application of data mining techniques for discovering interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications [6]. Usage data captures the identity and origin of web users along with their browsing behavior at a web site. Web usage mining tries to make sense of data generated by the web surfer's session or behavior [8]. Web usage mining itself can be classified further depending on the kind of usage data considered. First one is Web Server Data in which user logs are collected by the web server and typically include IP address, page reference and access time. Second is Application Server Data which track various kinds of business events and log them in application server logs. And third one is Application Level Data in which new kinds of events can be defined in an application, and logging can be turned on for them - generating histories of these specially defined vents.

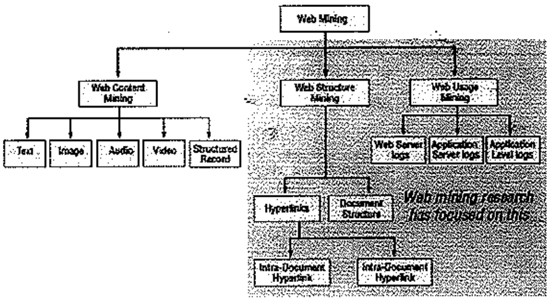


Figure1: Web Mining Taxonomy

## V. COMPARISON OF WCM, WSM AND WUM

A sufficient background to compare WCM, WSM and WUM has been created in the previous section. The following parameters are identified to compare them: View of data,

Type of data used for mining, Main Data, Representation, Method, Scope and Application Categories, Table1. gives the comparative study of these web mining types.

**Table 1:** Comparison of WCM, WSM and WUM

| **Specifications** | **Web Content Mining** | **Web Structure Mining** | **Web Usage Mining** |
|---|---|---|---|
| View of data | Structured, Semi Structured and Unstructured | Linking of structure | Interactive data |
| Type of data | Primary | Primary | Secondary |
| usedinmining Main data | Text document, Hypertext document | Link Structure | Server Logs, Browser Logs |
| Representation | Bag of Words, n-grams, Terms, phrases, Concepts or Ontology, Relational, Edge Labeled Graph | Graph | Relational Table, Graph |
| Method | Machine learning, Statistical Method, Proprietary Algorithm, Association Rules | Proprietary Algorithm | Machine Learning, Statistical Method |
| Tasks | It describes the discover you useful information from the web content/documents. | It tries to discover the model underlying the link structure of the web. | It tries to make sense of data generated by web surfer's session or behavior. |
| Scope | In R view of data these copies global while in DB view it is local. | Global | Global |
| Application Categories | Categorization ,Clustering, Finding Extraction Rules, Finding Pattern in text, User Modeling, Finding frequent Sub-Structures, Web site schema discovery. | Categorization, Clustering | Site Construction, Adaptation and Management, Marketing User Modeling |

**Table 2:** Comparison of Data Mining and Web Mining

| Comparison | Web Mining | Data Mining |
|---|---|---|
| **Scalability** | In this the search processing is so large, 10million job are in web server database | In this the search processing is not so large, about on Emily on job sare in the database |
| **Accessibility** | WebMiningisaccessingthedatapublicly.Inthiswedonot hidethedatawhichisaccessed in web d atabase. But we have to take permission to web log master to access the Data | Data Mining is accessing the data privately and only authorize user access the data that is present in the database |
| **Structure** | In this we take the structured, unstructured and semi structured data | Here we get the information from the explicit structure. In data mining we don't fetch the information from the wide database. |
| **Data** | Web mining works on on-linedata. | It works on off-line data |
| **Data Storage** | In this data is stored in web server database and server logs. | In this data is stored in data warehouse. |

### A. Discussion

The different parameters shown in Table1 are iscussed in detail below:

1) View of data: View represents the subset of data that is available on web. The data on the web can be in structured form, semi structured form or it may also be in the form of unstructured format. It is used to hide the complexity of data. In case of web structure mining the data is in the form of linking of structure. And in web usage mining the interaction of data with the user is needed.

2) Type of data used for mining: Web content mining and Web Structure Mining both uses primary data for the purpose of mining. Primary data is the data that is collected on source which has not been subjected to the pre- processing or any other manipulation. While Web Usage mining uses secondary data for the purpose of mining.

3) Main Data: The data that is actually used for the mining is the main data. Web content mining uses text document, hypertext document etc as the main data. Web structure mining uses link structure as the mining data while in web usage mining uses server logs, application server data and browser logs is uses as the main data.

4) Representation: How the data is shown on the web is the representation. Web Content Mining uses a large collection of words, various terms, semantics or ontology. It also uses relations to represent the data. Web structure mining uses graphs in which consists of web pages as nodes and hyper links as edge connecting related web pages. While web usage mining uses tables to represent the data and it also uses the graph t represents the data since it provides the interactivity of the data to the users.

5) Method: Various methods are used for mining the web data some of them are machine learning, statistical method, and proprietary algorithm and association rules. In machine learning various learning methods are used that is used to tell what the customer do and what he or she want. Various association rules are also used that is used to find the behavior of the user. For example "what is likely to happen in Delhi unit sales next or previous millions months.

6) Tasks: what is done by which type of mining is called its task. Web content mining describes the extraction of useful information from the web; web structure mining describes the link structure of the web i.e. how the data is linked from each other. While web usage mining describes the sense of data that is generated from the user's behavior [8].

7) Scope: There are two types of scope mainly Local scope and Global scope. Local scope spans individual web site while global scope spans the entire web. The scope of the Web content mining from the IR view and Web structure mining is global while the scope of the Web content mining from the DB view and Web usage mining is local.

8) Application categories: Where we use which technique is decided by the application category of various mining techniques. For ex clustering and categorization is used mainly in web content mining and web structure mining.

## VI. CONCLUSION

As the Web and its usage continues to grow, obviously grows the opportunity to analyze Web data and extract all manner of useful and interesting knowledge from it. Designing and maintaining web based information system such as web sites is a real challenge. An enormous amount of data is endlessly increasing on the web day by day. So it is much easier to find the inconsistent information than the well structured information so the study of web mining help a lot to analyze this huge collection of information that is available on web and it is also used to predict the behavior of user using various techniques. Hence a fraud and threat can be minimized.

## VII. FUTURE WORK

Here we provide a survey about the research in the area of Web mining's today structure and tomorrow view. We point out some confusion between data mining and web mining. Web data is growing at a significant rate. Web Mining is fertile area of research. Many Successful applications exist. We also suggest the subtask of web mining. Now the future scope of web mining is that we can also work for the process mining and try to combine usage mining with structure mining. We also go for the mining from cloud. Whenever we work on mining over cloud computing that time we hesitate for the cost but that come very less by cloud mining. So, we can say that cloud mining can seen as future of web mining and we hope this survey could be a very useful starting point for further research in future.

## REFERENCES

[1] O. Etzioni. *"The World-Wide Web: Quagmire or Gold Mine? Communications of the ACM"*, 39(11):65–68, 1996.

[2] R. Cooley, J. Srivastava, and B. Mobster. *"Web mining: Information and pattern discovery on the World Wide Web"*. In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 97), 1997.

[3] Srivastava, J., Cooley, R., Deshpande, M., And Tan, P-N. (2000).. *"Web usage mining: Discovery and applications of usage patterns from web data•h, SIGKDD Explorations"*, 1(2), 12-23.H. Poor, *"An Introduction to Signal Detection and Estimation"*. New York: Springer-Verlag, 1985,

[4] Rekha Jain and Dr G. N Purohit,. *"Page Ranking Algorithms for Web Mining"*. International Journal of Computer Applications (0975 . 8887 Volume 13. No.5, January 2011.

[5] Wang and Liu 1998, Moh, Lim and Ng 2000

[6] Srivastava, Cooley, *Deshpande*, and Tan 2000

[7] A. Houston, H. Chen, S. M. Hubbard, B. R. Schatz, T. D. Ng, R. R. Sewell, and K. M. Tolle. *"Medical data mining on the internet: Research on acancer information system"*. Artificial Intelligence Review, 13:437–446,1999.

[8] R. Cooley, B. Mobasher, and J. Srivastava. Web mining: *"Information and pattern discovery on the world wide web"*. In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 97),1997.

[9] Alexandra's Nanopoulos, Dimitris Katsaros and Yannis Manolopoulos *"Effective prediction of web-user accesses: A data mining approach,"* in Proc. Of the Workshop WEBKDD, 2001.

[10] Yang, Q. and Zhang, H. , *"Web-Log Mining for ng, IEEE Trans. Knowledge and Data Eng"*., 15(4), 2003,1050-1053.

[11] Jain Pei, Jiawei Han, Behzad Mortazavi_asl and Hua Zhu, *"Mining Access Patterns Efficiently from Web Logs"*, Pacific-Asia Conference on Knowledge Discovery and Data Mining(PAKDD'00), Kyoto, Japan, 2000, 396-407

[12] Manoj Manuja and Deepak Garg, " *Semantic web mining of Un-structured Data: Challenges and Opportunities"*, International Journal of Engineering, 5(3) ,2011, 268-276

[13] Jing Wang, Ying Liu, Lin Zhou, Yong Shi, and Xingquan Zhu, *"Pushing frequency constraint to utility MiningModel"*, ICCS Springer-Verlag Berlin Heidlberg, LNCS 4489, 2007, 685-692

[14] Miller, C.S. and Remington, *" R. W. Implications for information Architecture , Human Compter Interaction"*, Journal IEEE Web Intelligence, 2004, 19(3), 225-271.

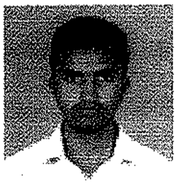
[15] Geeta.R.B, Shashikumar G. Totad & Prasad Reddy PVGD, *Optimizing User's Access To Web Pages*, International refereed Journal JooiJA, Transactions on World Wide Web-Spring, 2008, 8(1), 61-66.

[16] Garofalakis, *Web Site Optimization Using Page Popularity*, IEEE Internet Computing, 1999, 3(4), 22-29.

[17] Jia-ching Ying, Vincent S. Tseng, Philip S. Yu IEEE International Conference on Data Mining workshops IEEE Computer Society, 2009.

[18] Y.S.Lee, S.J Yen, and M.C.Hsiegh. A Lattice-*"Based Framework for Interactively and Incrementally Mining web traversal patterns: "*, International Journal of Web Information Systems, 2005. 197-207.

## AUTHORS' BIOGROPHY



**K.Vinodkumar** is an M.Phil scholar, Department of Computer Applications (MCA) in RVS College of Arts and Science, Coimbatore. He received his B.Sc., in 2010 in Bharathiar University and MCA in 2013 from Anna University, Chennai. He is pursuing his M.Phil. in the area of Data mining from Bharathiar University, Coimbatore. He got university rank in MCA.

**Kathiresan.V** is an Assistant Professor, Department of Computer Applications (MCA) in RVS College of Arts and Science, Coimbatore. He received his B.Sc., in 2003 and MCA in 2006 from Bharathiar University, Coimbatore. He obtained his M.Phil. in the area of Data mining from Periyar University, Salem in 2007. His research interest lies in the area of Data mining. He got Faculty Excellence Award from RVS College of Arts & Science for the Academic years 2007-08, 2008-09,2009-10, 2010-11, 2011-12 , 2012-13 and 2013-14 consecutively.