contents as well as such other content like images depends on the application requirements. WCM techniques can be utilized in many Web applications that aim to discover Web objects that having common characteristics or patterns such as group of Web pages that talking about similar topics (subjects), or similar Web images that includes certain objects like logos, watermarks or faces [4, 8]. Second Web Mining category is WSM, which concern with the process of discovering semantic features from Web pages such as the correlation among Web pages that belong either to similar or different Web sites. Hence, this category aimed to find group of Web pages or Web sites that relevant to each other based on structural links topologies, analyzing of this kind of data may require extracted massive links among Web pages, which produced fractal links when there is no correlation may exist among Web sites [6]. Finally, WUM also called Web log mining that attempts to extract, discover and analyzing interesting users' accesses, user transaction, clickstreams patterns and other associate data generated from user interaction with Web resources. Most Web development applications used Web Usage Data to analysis and extracted information about user profile, access patterns, familiar pages' contents, actions, and others, which used by many World largest companies such Yahoo, MSN, Amazon and others to discover their users' access patterns [7].

TABLE I
WEB MINING TECHNIQUES AND APPLICATION

Web Mining Taxonomy				
Approach	**Web Content Mining**		**Web Structure Mining**	**Web Usage Mining**
	IR view	***DB view***		
Data view	- Structured - Unstructured	Semi-structured	Link structure	Interactivity
Data source	- Web documents (text, image, video, Audio)	Hyper-text documents	Link structure (inter links / intra links)	- Server log - Proxy log - Client log
Representation Model	- Text based (VSM, n-gram) - Phrase based, concepts. - relational	- Labelled graph - Relational	- Graph	- Relational table - graph
DM techniques	- Machine learning - Statistical	- Proprietary algorithms - Association Rule	- Proprietary algorithms	- Machine learning - Statistical - Association Rule
Application Area	- Classification - Clustering	- Sub-set structure - Web site Schema	- Classification - Clustering	- Classification - Clustering - Association Rule

II. MATERIAL AND METHOD

WCM refer to mining, extracting and integrate of valuable information (knowledge) from Web pages' content, which including several techniques to mining different data types, the following figure 1 illustrate the current and novel techniques that used to analyzing the Web contents [4, 12, 8].

A. Unstructured Data Techniques

Most data that available on the Web is in unstructured form, unstructured Web mining techniques illustrates as follow:

1) Information Extraction (IE): This technique is very interest when there is a large volume of accumulated text. This technique can provide KDD by transform unstructured data form to structured form then mined information by using several rules, IE that made incorrect prediction on data are discarded [11].

2) Topic Tracking: The process of finding Web documents that related to user query by identifying the Web pages that related to certain topics, then using hyperlinks to identify group of Web pages that relevant to a specific topic [5, 12].

3) Summarization: Is a technique that used to reduce the length of Web documents and give decision to the user if should read these documents (pages) or not instead of reading the first paragraph to know if the Web document is related or not to his interest [5].

4) Categorization (Classification): Classification is the process of assigning class label to the Web contents from set of pre-defined classes (labels) in dataset, classification of the Web contents can be dividing into [8, 12]:

- Web Page Classification (Categorization)

Web page classification is a process of assigning class or category to the Web page from set of prior classes or categories. The different between Web categorization and Topic tracking is the last one is concerned with the content-oriented topics [5]. Web page categorization is differing from traditional text document categorization process by the following aspects, first, categorization of traditional text document is typically implemented on structured and consistence style while Web page do not have this attribute. Second, Web pages mostly exist in unstructured HTML forms and includes such other attributes that do not exist in text documents such as keyword, title, head and description tags [13].

- Web Sites Classification

Web site rather than Web page also can be classifying; many approaches used for Web sites classification, one approach uses Web site's home page contents based classification [15], other approach is used HTML tags to classifying Web sites such as classifying sites into industry categories as in [16], while another is used link structure attributes based classification [17]. Web page classification may support Web site classification by knowing the topic of Web page we know also the global Web site topics [18].

5) Clustering: Clustering is a technique that used to group similar objects, Web environments include several Web contents, which clustered based on certain characteristics or parameters for example Web pages clustering. Clustering Web can be one of the following types [12, 14].

- Web Page (Document) Clustering

Web pages are grouping based on such related contents, this information is useful in search engines and information retrieval approaches, which increased the accessibility and develop content, based delivering applications [14].

- Web Objects Clustering

Clustering of Web objects establish a method for grouping relative content to serve user queries, objects may include text files, images, videos and sound tracks [12].

- Web Sites Clustering

Web sites clustering is a process that aim to group similar Web sites that having similar characteristics. There are many critical challenges to implement this task, which includes extracting of the textual content from several Web pages is a complex task and requires many pre-processing steps, second, mining multimedia content such as images, audios and videos need novel techniques that do not implemented in large computation time [13, 15].

6) Information Visualization: Web contents can be better comprehended by using visualization tools. Thus, there is a necessity to improve a tool that able to providing a graphical representation of Web objects. These tools are able to visualizing significant information such as Web usage time, users' clickstreams, Web sites relationships, Web pages' correlations, users access patterns…etc. in conceptually, there are many visualization tools used with Web contents such as STATISTICA, NCSS, Ggobi and T-SNE, where user can visualize large contents in pictorial forms like statistics pars, histograms, scatter plot and many other [8, 12, 15].

B. Structured Data Techniques

Applying Data Mining techniques to extracted information from structured data as in the Web pages, thus structured data in Web pages comes in the forms of tables, list and tree. Structured data by compared to unstructured forms it's easy to extract, the techniques used with structured data types described as follow [19].

1) Web Crawler: Also known as spiders or robots, crawler are the programs that automatically traverse through hyperlink structure and download correlated Web pages. There are many applications for crawling software such as Web site monitoring, Web pages' categorization, Web content updating and business intelligence [20].

2) Wrapper Generation: Wrapper Generation can be defined as the process of ranked Web pages based on Web rank value for retrieving related Web pages by search engines according to the query that made by users through their Web browsers [22, 23].

3) Web Page Content Mining (WPCM): WPCM is a mining process of the entire content of Web pages by search engines to ranked Web pages through searching process, this process includes extracting structured data, then list the result based on its correlation [14, 22].

C. Semi-Structured Data Techniques

Including the techniques that used to extract data combined from heterogeneous sources such as Web page in database, books and Author's which have been written in semi-structured forms, these techniques includes the following [12, 14, 22]:

1) Object Exchange Model (OEM): Defined as the process of extraction useful information from semi-structured contents, then segment similar content into coherent groups [14].

2) Top-Down Extraction: Defined as the process of extraction complex objects from different Web resources and decompose them till tiny objects have been extracted and collected [16].

3) Web Data Extraction Language: This technique used to convert semi-structured Web contents into structured forms, then deliver it to the end-user, such as RDF (Resource Description Framework), XMDP (XHTML Metadata Profiles) and JSON-LD (JavaScript Object Notation for Linked Data) [21].

D. Multimedia Data Techniques

Multimedia techniques are the processes of extracting and discovering knowledge from multimedia dataset like text, images, audios, videos and combination of them from large multimedia database that are not ordinarily accessible by basic user queries. Multimedia techniques involve some Data Mining techniques such as classification, clustering, sequence patterns mining, association rule and standard statistics [22]. The main goals behind Multimedia Data Mining System (MDM) are similarity search in multimedia data, indices construction and retrieval implementation, description-based retrieval by using (keywords, tags, caption, time stamp and size), content-based retrieval such (color, histogram, texture, shape, wavelet transform). Multimedia mining involving two basic steps, which are extracting underling features from interesting data and selecting multimedia mining method to identify desired contents. Multimedia techniques categorized to the following techniques [23]:

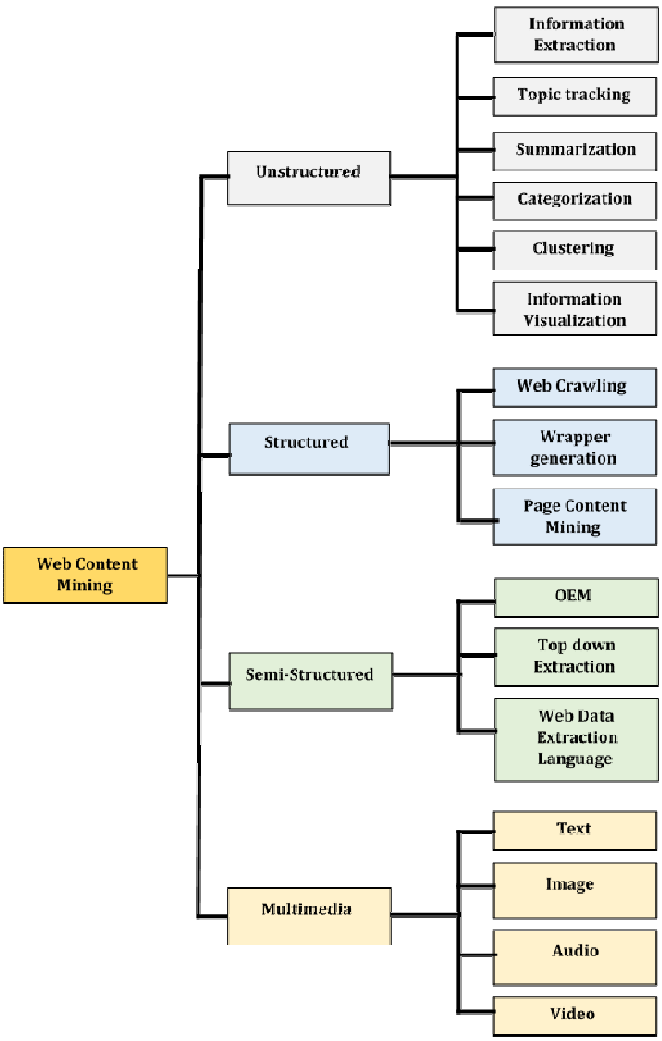

Fig.1. Web Content Mining Techniques and Applications

1) Text Mining: Text mining refers to the process of applying Data Mining techniques that to extract significant text portions (blocks) from unstructured Web documents. Bag of Words (BOW) is a common model used in Web mining to represent the absence or presence of textual features (e.g. sentences or words), this model also called Vector Space Model (VSM) [9].

2) Image Mining: The technique that used to discover image patterns or useful information from large collection of images called image mining. Image mining may conclude several approaches such as digital images processing, image understanding, multimedia databases, Artificial Intelligent…etc., there are many proposed methods that focused on processing the images itself to extract the desired features such as texture analysis, line detection, smoothing, color histogram and many others to solve the task of image analysis [22].

3) Audio Mining: This technique is concern with the process of analyzing and searching through an audio content, it is generally use with the field of speech recognition [14].

4) Video Mining: Video mining involves the tasks of digital video processing which includes automatic segmentation, indexing, content-based retrieval and classification of visual objects [8].

III. Results and Discussion

The main goal behind WSM is to extract previously unknown relationships among the Web pages that belong to single or many Web sites. There are two different approaches in WSM, which are link topology mining and link URL mining; both approaches used a different raw data and methods. Topology mining percept the Web as a graph in such a way the web pages represented as nodes and the edges are the hyperlinks among these pages [7]. URL mining conjunction with link topology for the source and targets pages to construct more accurate link patterns model. WSM can be used in future works to analyzing social networks and identifying users' communities that having common information sources [6], application upon WSM are illustrates as follow:

A. Clustering

Web includes several objects that exist mostly with no integrated structure; the objects in the WWW are the Web pages, which are linkage to other pages through the links [8]. The main goal of clustering technique is to group similar pages based on the kind of structure information used which includes: Hyperlinks, Document structure and Link analysis. Therefore, clustering enables of connected Web pages to establish relationship of other related pages and allows users to access the desired information through keyword association and extracted contents [21, 22].

B. Classification

Classification is a supervised Data Mining techniques that aim to assign class property from set of predefined set of classes, in Web data there two types of classification which are [14, 15]:

1) Link-Based Classification: Link (hyperlink) based classification is the most recent upgrading technique in Web mining, the main goal behind it is to predict the category of the Web page based on the links attributes (e.g. Links among Web pages, Anchor text and other HTML tags) [17, 19]. Hyperlinks is a structural portion that connect different location in Web pages to other location either within the same page or other pages that belongs to single or many Web sites. Link (hyperlink) structure classified into two types, which are inter-document hyperlinks and intra-document hyperlinks. Inter-document hyperlink defined as the hyperlink that connect different Web pages while intra-document hyperlink connects different parts within the same Web page [28].

2) Content-Based Classification: Content based classification aims to classifying Web page based on the link (hyperlink) contents (anchor text of the link). Hence, every Web page assigned to class based on the words that appears in their links, this approach is required iterative technique for assigning the labels, due to class of Web pages may potentially changes through the links that scattered on the Web page itself [14].

C. Retrieval

Information that containing in the hyperlinks plays an important role for retrieving the results in the search engines, where the anchor text portions (text that appear in hyperlinks) of predecessor Web pages are already indexed by World Wide Web Worms. There are two types could be retrieved based on user query: authorities' pages and hubs pages [14, 30]. Every Web page assign two scores, one is called authority score and the other its hubs score by using an algorithm called HITS (hyperlink-induced topic search) which comes to solve the problems of Web search engines [30]. Authority's pages are the pages that containing significant information about the query topic, while the pages that points to many authority Web pages is called hub pages and its useful resources in the Web. Hence, the scores determined which page is a good hub page if it is containing pointers to many good authorities, and good authority page if many good hubs pages point it and pages are ranked based on score values in Web search engines [14, 16, 30].

D. Web Usage Mining

WUM also called web log mining that attempt to extract, discover and analysis interesting users' accesses, user transaction, clickstreams patterns and other associate data generated from user interaction with Web resources and stored it in standard text file format called log file that reside on the Web server itself. Most web development applications used Web usage data to analysis and extracted information about user profile, access patterns for the pages' contents, actions, and others and it used by many World largest companies such Yahoo, MSN, Amazon and others to collect data from Web log access to discover their users' access patterns [7, 9].

In WUM, data can be collected from different levels or obtain from an organization as dataset such as NASA, the data collected differ in terms of content structure, sources, type of information available, segment methods and implementation method. Data source levels of Web Usage Data categorized into: Client level data, Proxy level data and Server level data [10, 11], figure 2 shows the architecture level of Web data sources.

- Client Level Log

This kind of log file reside in client's browser window which include information related to single user actions towards single or many Web sites browsing behavior [5]. Cookies used to store status information of Web site such passwords and some related information like passwords, and this information stored in client machine and can recognize this user from others due its contain information about user's browsing operating system, this kind of log file in many Web applications is not considered due to client cookies may disable by user for security and privacy issues and collecting information from all users is a practically difficult task [27].

- Proxy Level Log

Proxy Log act as an intermediate stage that take HTTP requests from whole users and passing it to Web server, then returns the results by the Web server and passing it again to the submitted users. Web proxy log file can record all the requests that made by users' communities that access to the Internet through Internet Service Providers (ISPs), it's possible to identify host machine name making that request along with other underling information [1, 14]. The drawbacks of Proxy level are the construction Proxy server is a difficult task and it required advanced networking programmers, and request interception is limited [28, 29].

- Server Level Log

Server log file provide most accurate and complete usage data when users interact with various Web sites. Log data is primary used in WUM which contain access log data and application log data. Pre-processing of web data is a practical challenge in many Web applications due to the following reasons: the scale of Web data exceeds any conventional database, log data stored in Web server log file is in standard text file format and comes in different formats, inferring cached pages' references and clickstream patterns and its relationship to other related data all these reasons make preprocessing is often most time consuming and computationally insensitive step [3, 7].

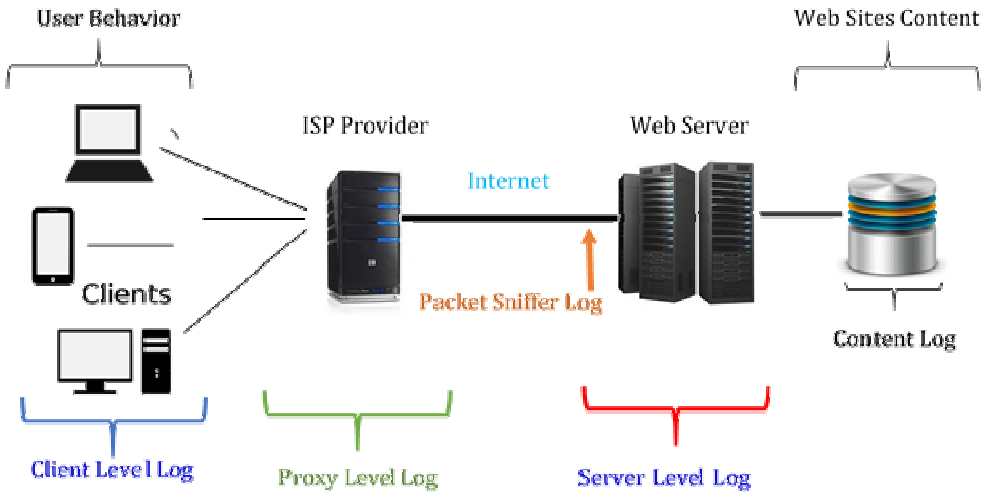

Fig.2. High level architecture of different Web services

WUM often required to use special and heuristic algorithms which are not commonly used in other domains, many research work focused only on pre-processing and integration data sources for varies analysis, successfully employment of Data Mining techniques in Web Usage Data is highly correlated with the pre-processing tasks [7].

There are different WUM techniques applied to extract knowledge from log data described as follow:

1) Classification: There are many algorithms can be used for classifying users into set of pre-defined classes, this process includes training a classifier algorithm along with training data then used for classify tested part [8].

2) Clustering: The technique that groups' users with similar characteristics in Web Usage Data is called clustering, there are two types of clustering can be addressed here which are: usage clustering and pages clustering, usage clustering tends to group Web users with similar navigation patterns, while page clustering tends to group pages that having related contents [10].

3) Statistical Analysis: One of the common method to extract knowledge about the Web is Statistics measures, which present analyzing about numerous entries and contents. The statistical analysis techniques are most powerful tools to extract knowledge about users' activities towards several Web sites contents. The extracted statistical reports can help for improving the performance of Web sites, providing support for marketing decision, enhanced security system, predict user activities and many other useful application [4].

4) Association Rule: Association Rule used to predict the next event of user by analysis the total transaction to the number of URLs in the Web log data that made by that user. By applying different Association Rule mining algorithm, we can predict what are the Web sites visited together, Web pages accessed frequently and similar interesting content requested. Discovering of such rules help also in recognizing the structure designing of the Web sites, thus we can know the confidence of requested contents such as items, pages or products by analysis the sequential path of user accesses in log file data [28].

E. Future Direction

Web Mining become an integral field to many web development applications while a lot of organizations content reside on the web are increased enormously. Web mining is a novel extension of Data Mining techniques that applied on the data types that reside on the Web which exist in different forms and structures. Therefore, there is persistent need to cover and provides rigorous solutions for many critical problems such as controlling, monitoring, perception, and knowledge representation of Web data generated either by users' communities or dispatching of database system to web services. Our future work focused on combination more than Web mining technique to provide a better perception for such complex Web data and analyzing the methods that follow to discover the significant patterns from the Web, such as analyzing Web usage data and recent techniques used for discovering usage patterns from it, and analyzing the significant parts in Web pages' content to be used for significant application such as recommendation system or Web pages' classification / clustering.

IV. CONCLUSIONS

In this paper, we survey various Web mining techniques that used by a lot of Web application recently. We had also reviewed a comparison among Web mining categories based on significant approaches used currently by most of research works. Since, Web environment is a huge area and there are a lot of work to do in future, we hope this paper could be providing a good starting points to knowing current Data Mining techniques that applied upon different Web data and also help to identifying opportunities for forthcoming research works by understanding the nature of the data that reside on different Web resources.

REFERENCES

[1] Bing Liu, "Web Data Mining, Exploring Hyperlinks, Contents and Usage data", 2nd edition, Springer New York, ISBN: 9783642194597, PP: 1-14, 2011.

[2] S. G. Langhnoja, M. P. Barot, D. B. Mehta, "Web Usage Mining ton Discover Visitor Group with Common Behavior using DBSCAN Clustering Algorithm", International Journal of Engineering and Innovative Technology, Vol.2, No.7, pp. 169-173, 2013.

[3] Emrah Donmez, Alper Ozcan, "Time Based Discovering of Web User Patterns to Optimize Web Sites and Hyperlinks", International Journal of Advanced Computational Engineering and Networking, Vol.3, No.2, pp. 14-20, 2015.

[4] Shyam N. Kumar, "World towards Advance Web Mining: A Review", American Journal of Systems and Software, Vol.3, No.8, pp. 44-61, 2015.

[5] J. Srivastava, P. Desilkan, V. Kumar, "Web Mining- Concepts, Application and Research Directories", Foundation and Advances in Data Mining, Wesley W. Chu and T. Y. Lin, Springer-Verlag, pp. 275-307, ISBN: 9783540250579, 2005.

[6] Neha Sharma, "A Review on Analysis to Improve Performance of Website", International Journal of Science and Research, Vol.6, No.6, pp. 2453-2455, 2017.

[7] Bamshad Mobasher, Olfa Nasraoui, "Web Usage Mining ", in Web Data Mining, Exploring Hyperlinks, Contents and Usage data, Bing Liu, 2nd edition, Springer New York, ISBN: 9783642194597, PP: 527-603, 2011.

[8] Anurag kumar, Ravi Kumar Singh., " A Study on Web Content Mining", International Journal of Engineering and Computer Science, Vol.6, No.1, PP: 20003-20006, 2017.

[9] Charu C. Aggarwal, "Mining Text Data ", in Data Mining the Text book, C. C. Aggaral, Springer Switzerlands, ISBN: 9783319141411, PP: 429-456, 2015.

[10] Tanveer Kaur Dewgun, Pushpraj Singh Chauhan, "A Survey on Web Usage Mining: Process, Techniques and Applications", International Journal of Engineering Research and Technology, Vol.4, No.4, pp. 1013-1015, 2015.

[11] Faustina Johnson, Santosh Kumar Gupta, "Web Content Mining Techniques: A Survey", International Journal of Computer Applications, Vol.47, No.11, pp. 44-50, 2012.

[12] Monika Yadav, Pradeep Mittal, "Web Mining: An Introduction", International Journal of Computer Science and Software Engineering, Vol.3, No.3, pp. 683-688, 2013.

[13] Shanthi S, " Survey on Web Usage Mining using Association Rule Mining", International Journal of Innovative Computer Science & Engineering, Vol.4, No.3, pp. 65-67, 2017.

[14] Athena Vakali, George Pallis, Lefteris Angelis, "Clustering Web Information Sources"; In Web Data management practices: Emerging Techniques and Technologies, IDEA group publishing, pp. 34–55, ISBN: 1599042282, 2007.

[15] Xiaoguang QI, Brian D. Davision, 2009. "Web Page Classification: Features and Algorithms", ACM comput. Survey 41, Article 12, http://doi.acm.org/10.1145/ 1459352. 1459357 (Accessed on July 7, 2016).

[16] John M. Pierre,2001."On the Automated Classification of Web Sites, Link¨oping Electronic Articles in Computer and Information Science, Vol.6, http://www.ep.liu.se/ea/cis/2001/000/, (Accessed on July 7, 2016).

[17] Einat Amitay, David Carmel, Adam Darlow, Ronny Lempel, Aya Soffer, " The Connectivity Sonar: Detecting Site Functionality by Structural Patterns", In Proceedings of the 14th ACM Conference on Hypertext and Hypermedia (HYPERTEXT). ACM Press, New York, NY, pp.38–47, 2003.

[18] Martin Ester, Hans-Peter Kriegel, Matthias Schuber, "Web Site Mining: A new way to spot Competitors, Customers and Suppliers in the World Wide Web", In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). ACM Press, New York, NY, pp. 249–258, 2002.

[19] Shipra Saini, Hari Mohan Pandey, "Review on Web Content Mining Techniques", International Journal of Computer Applications, Vol.118, No.18, pp. 33-36, 2015.

[20] Filippo Menczer, "Web Crawling", in Web Data Mining, Exploring Hyperlinks, Contents and Usage data, Bing Liu, 2nd edition, Springer New York, ISBN: 9783642194597, PP: 311-362, 2011.

[21] Leslie F. Sikos, "Mastering Structured Data on the Semantic Web from HTML5 Microdata to Linked Open Data", Apress, New York, ISBN-13 (electronic): 9781484210499, PP: 256, 2015.

[22] Sarla More, Durgesh K. Mishra, "Multimedia Data Mining: A Survey", PRATIBHA, International Journal of Science, Spirituality, Business and Technology, Vol.1, No.1, pp. 49-55, 2012.

[23] S.Vijayarani, Ms. A.Sakila, "Multimedia Mining Research – An Overview, International Journal of Computer Graphics & Animation, Vol.5, No.1, pp. 69-77, 2015.

[24] Miguel G. C. Júnior, Zhiguo Gong, "Web Structure Mining: An Introduction, Proceedings of the International Conference on Information Acquisition, IEEE, Hong Kong and Macau, China, pp. 590-595, 2005.

[25] Mike Thelwall, "Data Cleaning and Validation for Multiple Site Link Structure Analysis ", In Web Mining: Application and Techniques, Anthony Scime, IDEA group publishing, USA, UK, pp. 208-227, ISBN: 1591404169, 2005.

[26] Bamshad Mobasher, Olfa Nasraoui, "Web Usage Mining ", in Web Data Mining, Exploring Hyperlinks, Contents and Usage data, Bing Liu, 2nd edition, Springer New York, ISBN: 9783642194597, PP: 527-603, 2011.

[27] Charu C. Aggarwal, "Mining Web Data ", in Data Mining the Text book, C. C. Aggarwal, Springer Switzerlands, ISBN: 9783319141411, PP: 589-617, 2015.

[28] Shaily G.Langhnoja, Mehul P. Barot, Darshak B. Mehta, "Web Usage Mining using Association Rule Mining on Clustering Data for Pattern Discovery", International Journal of Data Mining techniques and Application, Vol.2, No.1, pp. 141-150, 2013.

[29] Wen-Chen Hu, Xuli Zong, Chung-wei Lee, Jyh-haw Yeh, "World Wide Web Usage Mining Systems and Technologies, Journal of Systematic, Cybernetic and Informatics, Vol.1, No.4, pp. 53-59, 2014.

[30] Rashmi Sharma, Kamaljit Kaur, " Review of Web Structure Mining Techniques using Clustering and Ranking Algorithms", International Journal of Research in Computer and Communication Technology, Vol.3, No.6, pp. 663-668, 2014.