A Systematic Review Web Content Mining Tools and its Applications

Systematic Review Web Content Mining Tools

Manjunath Pujar[1], Monica R Mundada[2]
Department of Computer Science
M. S. Ramaiah Institute of Technology, Bangalore, Karnataka, India - 560054

***Abstract*—In recent years, the emergence of WWW (World Wide Web) led to the accumulation of huge amount of information and data. Hence the web is found to consist of unstructured and structured information that impacts the day to day life of the society. Because of such availability of huge information, utilization of the required information becomes more challenging. This paper provided a comprehensive survey on the current situation and recent trends on web content mining (WCM) and its applications thereby contributing to the enhancement of the upcoming research in WCM. The paper focused mainly on the mining and retrieval techniques, various WCM approaches, challenges and process of information retrieval and information extraction. The paper describes the four major tasks of web content mining that is information retrieval, information extraction, generalization and validation in detail. WCM concentrates on orchestrating, sorting, classifying, collecting, congregating of web data and provide the improved data which can be easily accessed by the users. Web content mining tools were needed to scan text, images and HTML documents and provide results to the search engine. It guides the search engine to provide better productive results for every search based on their importance. The paper also analysed different web content mining tools for the extraction of relevant information from the corresponding web page.**

Keywords*—Web content mining; web structure mining; web usage mining; data mining; information retrieval; information extraction

I. Introduction

There exist several under-research fields, such as relevance ranking of webpage, knowledge in web documents, etc. in the Web Content Mining. The extraction of useful patterns and information from the page content, web hyperlink structure and usage data with the employment of data mining and artificial intelligence methods is termed as web content mining. Few techniques such as association rule mining, classification analysis, support vector machine, clustering analysis are utilized for the purpose. The main contribution of the review is to frame a semi structured and comprehensive overview of the problems, methods and solutions of the prevailing web content extraction problems [1].

Enormous growth of data leads to numerous creation of web pages for analyzing and mining of useful data but it is practically challengeable. WWW includes billions of millions of interlinked web pages created by billions of millions of authors on the world. Web page is kind of document utilized to view WWW with use of web browser [2]. Massive source of data in form of unstructured and structured data is added to the web daily, which makes data extraction complex. Technique for efficient data mining from the web is imperatively prominent requirement for the user. Data mining is widely utilized technique in government security, science explores and business [3, 4]. Web mining is an application of data mining method which is semi-structured or unstructured data, automatically find and extract most useful, unknown information from the web [5]. Few applications of web mining is: web communities, web market places, artificial intelligence, business, e-commerce, network management, information retrieval, search engines, web search and web design. Process of web mining consist of 4 steps such as: finding of resources, selection of data, pre-processing generalization and examination of resource findings. Resource finding is process utilized to mine data either from offline or online resources. In selection of data and pre-processing step, information which is retrieved from web resources is selected and pre-processed automatically. During generalization step, machine learning and data mining methods were utilized to find common patterns from every websites. In analysis step, extracted data undergoes validation and interpretation.

This study includes six sections such as: Section 1 gives introduction to data and web mining. Section 2 explores categories of web mining. Section 3 and Section 4 elaborates tools and techniques related to content mining. Section 5 gives comparison of web content mining tools and Section 6 concludes this survey with future work.

II. Categories of Web Mining

Web mining is categorized into three major types (3):

- WSM (Web Structure Mining).
- WUM (Web Usage Mining).
- WCM (Web Content Mining).

A. WSM

WSM is a study of data related to structure of specific website [6]. It includes web graph which consist of web reports or web pages as nodes and hyperlink are consider as edges which links two connected pages. WSM is process of mining web graph patterns such as complete graph, social choice, co-citation, etc. Sited page is ranked by varied points and web page is selected in order to include within page group. This

kind of mining is performed at page level or in-between page level [5]. This mining determines the structure knowledge from the web link structure. Concentrating on web link structure is big challenge for this kind of mining. The main use of this mining is to examine the web pages and arrange in structurally manner, Fig. 1.

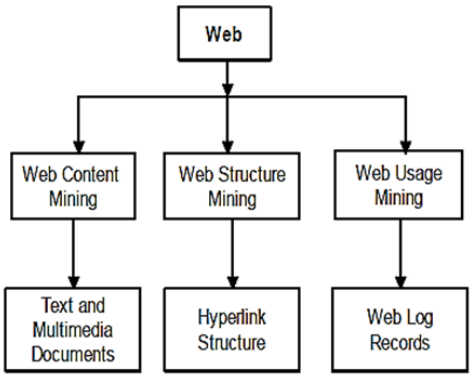

Fig. 1. Classification of Web Mining [7].

B. WUM

WUM also called as web log mining mainly utilized to examine users behavior on web. This mining concentrates on methods which predicts user knowledge when they are in online. Primarily it concentrates to track behavior of the user based on their personalized and generalized web usage [8]. According to generalized tracking extraction of knowledge is done related to user's web page history. Whereas in personalized tracking, extraction of knowledge is done from web interaction of users. Sources of this mining are logs of proxy server, server logs, bowser logs, and client logs. Main purpose of this mining is data pre-processing which includes data identification, cleaning and developing user sessions, extracting information of particular web page and formatting that particular data. Once pre-processing of data gets completed then data mining algorithm is used to detect behavior of the user.

C. WCM

It is a process of mining necessary and beneficial data from the web pages. It's mostly linked with text mining since most of the web content consist of text. It also concentrates on content of web page like images, text and some other media files which are attached. Here pre-processing is carried out on content of web page. WCM is utilized in various web applications with intension to identify web objects which have common patterns or characteristics [9, 10]. It is naturally semi-structure format of web. It has two kids: one type directly extracts document's content and another type enhance search of content with tools like search engine. WCM is utilized to analyze collected data with help of web spider and search engine. Generally utilized technology in this technique are NLP (Natural Language Processing) and IR (Information Retrieval). Two methodologies utilized in ECM are agent-based and database approach. Where Database method will retrieve semi-structured data from the web document. Agent-based method use three types of agents such as: intelligent search agent, information filtering classifying agent, customized web agent. Where customized web agent tries to determine web page on user profile. Information filtering classifying agent minimize effort and time of users in finding relevant document. Intelligent search agent automatically finds information related to the query of users. Various categories of WCM is discussed in below section.

III. Various Techniques of WCM

WCM were used to extract information from web pages [5]. Various techniques of WCM is illustrated in Fig. 2.

- Unstructured Content mining.
- Structure data mining.
- Semi-Organized data mining.
- Multimedia data mining.

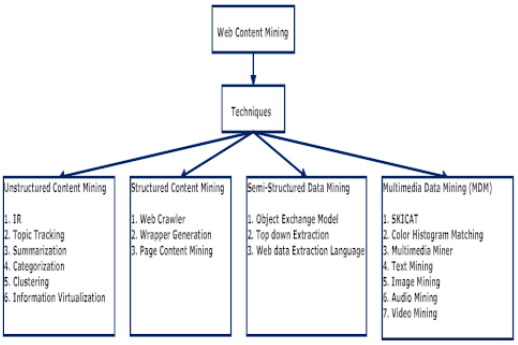

Fig. 2. Techniques of Web Content Mining [13].

A. Unstructured Content Mining.

Mostly web content will be unstructured data format. Using data mining methods with unstructured data is referred as KDT (Knowledge Discovery in Text) [10]. Hence text mining is consider as instance of web content mining. For effective result, pre-processing step must be carried out with help of NP, IR or text categorization techniques [11].

1) Units: Information can be mined from unstructured data with use of pattern co-ordination. IR tracks the expressions and keywords and also finds association of keywords within the content [12]. This method is suitable for large size text. IR is basic for all technologies which is utilized in un-organizing mining and also finds KDD, because IR converts unorganized data into more organized data structure. Primarily, data is mined from the mined data then various kinds of rules were utilized to find the lost information.

2) Topic tracking: This technology analyzes the document based on user view and predicts the document based on their interest. It is a process determining web document related to query of the user by finding the web pages which is related to user topics and also use hyperlinks to find set of web pages which is related to user topic.

3) Summarization: It is utilized to minimize document length by controlling the main points. It supports the users for taking decision whether it can be read or not. Time taken for summarization of document is minimum when compared to time taken to read the 1st paragraph by the users. One of the

risk with summarization was to teach the software to examine semantics and to understand the meaning. This software statistically balances the sentence and mine the needed sentence from documents. Summarization tool is used to understand key point of the documents. This tool provides freedom to user to choose the percentage of text to be mined as summary. It also work with categorization and topic tracking to summarize the document. Microsoft auto summarize word are example of summarization.

4) Categorization: This technique position the document in pre-defined set of set. It counts total number of words within document based on this main topic is decided. Rank is awarded to document based on the topic. Document which gives major preference to particular topic are considered first. This technique supports business and industries. Categorization is divided into two types.

5) Web page categorization: It is process of allocating category or class to web page from group of prior categories or classes [14]. This categorization differs from conventional document categorization process by the features like, primarily traditional text document which is employed on consistent style and structure format where web page won't have attribute. Secondly, generally web page exist in unstructured HTML format and has attributes like description, head, tittle and keyword [15].

6) Web site categorization: Various techniques were used to classify web site. One of the technique is web site home page content dependent classification and another technique is HTML tag utilized to categories websites as performed in 16 and another technique utilize link structure attribute dependent classification. Website classification is also supported by webpage classification 18.

7) Clustering: This technique is utilized to cluster similar documents. Here grouping was performed depending on fly not on predefined topics. Same document appears in various group. Because of this, useful document won't be deleted from search the result. This technique supports the user in easy selection of topic of user interest. This technique is utilized in MIS (Management Information Systems). Clustering can be done in different ways such as:

a) Web page Clustering: This technique group's webpages based on related content, this is useful for IR approaches and search engines, which enhance accessibility and create content depending on delivery applications.

b) Web Object Clustering: IT groups objects such as sound tracks, video, audio, images and text according to the user queries.

c) Web Site Clustering: This technique groups similar websites which have similar characteristics. There are numerous challenges with implementation such as: mining textual content from various webpages is tedious process and needs several pre-processing steps, and then extraction of multimedia content like video, audio and image requires new technique which is complex to implement.

8) Information virtualization: Web content can be compressed with use of virtualization tools. So there is need to enhance tool in a way to provide graphical representation of web objects. These tools visualize information, like user access pattern, webpage correlation, website relationship, users click stream and web usage time, etc. There are numerous visualization tools like T-SNE, Ggobi, NCSS, STATISTICA, these tools provide the content in form of scatter plots, histogram, etc.

B. Structured Content Mining

This type of mining is simpler compared to unstructured data mining. In webpage, host page are called as structured data. This process have three techniques, such as: page content mining, web crawler and web generation. Process of mining structured data is referred as wrapper. Data retrieved from the database with use of pre-defined structure in webpage called as structured data [16]. This technique supports user to organize data from various sources [11].

1) Web crawler: It's also called as web spider is an intelligent software program used to search over the web for particular information. Search engine with support of web crawler provide fast search result. This is performed by duplication browsed data from webpages for future action and the downloaded pages are indexed [17]. It also automatically update the indexes and web content. For extraction of information, web crawlers utilize techniques such as soft computing techniques, PSO, genetic algorithm and breath-first search. And also traverse through hyperlink to mine knowledge.

2) Wrapper generation: Wrappers serves Meta information like statistics, index links, and source domain. In this technique, depending on source capacity information is provided. Wrapper mine content from particular source of data to convert it to link structured manner [18]. There are two ways to perform wrapper generation they are: automatic and induction data extraction. For induction type, supervised learning is utilized to extract information from trained program given by users. Automatic data extraction is all ways possible since most of the web data is in generic form.

3) Page content mining: This process categorizes web pages. This techniques functions based on page ranking provided by standard search engines.

4) Semi-Structured data mining: It is not grammatical or full text. Semi-structure data won't have pre-define structure, it will be in hierarchical in nature. There are various methods to mine semi structure data like NP, ontology, wrapper generation. To mine such kind of data there is need for general technique to develop particular grammar in a way of briefing to extract surrounding piece of data.

5) Object Exchange Model (OEM): Relevant data is mined from semi-structure and is grouped in form of useful information and preserved in OEM. This supports user for easy and clear understanding of information structure which exist in web. This model is self-describing so there is no necessary for description of object structure in advance [7].

6) Top down extraction: This method supports to mine difficult items from off-sources of web and break those difficult items to less complicate items. This process continues till atomic elements were separated.

7) Web data extraction language: This technique converts to structured data from web data which was given to end-users. This technique stores data in table format.

C. Multimedia Data Mining (MDM)

It is a process of determining patterns from media data like image, text, audio and video which are not accessible with use of queries [7]. MDM involves with standard statistics, association rule, sequence pattern mining, clustering and classification. Intention of MDM is to find indices construction, multimedia data, retrieval implementation and description dependent retrieval with use of size, time stamp, caption, tags and keywords, and content dependent retrieval with use of wavelet transform, shape, texture, histogram, color. MDM involves in two steps like: mining features from data and choosing multimedia mining techniques to find preferred contents. MDM is categorized as follows [19]:

1) SKICAT: It is astronomical cataloging and data analysis system which develop digital catalog of sky objects. It utilizes machine learning methods to translate object into human classes. It combines methods for data classification and image processing which supports to categorize huge classification sets.

2) Color histogram matching: This technique comprise of smoothing and shading histogram equations. Equation tries to find relationship among parts of color. Equalization is challengeable task due to existence of unwanted shortages in adjusted pictures. Smoothing is used to resolve this issue [7].

3) Multimedia miner: It consist of 4 steps they are: In step 1, image excavator for mining videos and images. In step 2, preprocessor for mining image features and preserved in database. In step 3, search kernel is utilized for matching queries with videos and images which exist in database. In step 4, discovery modules extract image information to trace patterns in image.

4) Text mining: It is process of using data mining methods to mine text portions from unstructured web documents. BOW (Bag of Words) is general model utilized in web mining to indicate presence or absence of textual features. BOW is also referred as VSM (Vector Space Model) [20].

5) Image mining: It is a process of finding image pattern from huge collection of images. It concentrates on image processing to mine desired features like color histogram, smoothing, line detection, texture analysis to solve analysis of image.

6) Audio mining: It is a process of examining and searching over audio content, particularly utilized with speech recognition [19].

7) Video mining: It is a process of digital video processing which involves in classification of visual objects, content-dependent retrieval, indexing and automatic segmentation [10].

This method is utilized to process video content and also predicts transitions among two frames like former analysis multimedia content.

IV. Various Approaches of WCM

WCM is process of mining needed knowledge from web document. The content may be video, audio, image and text. In WCM issues are related to IR and visualization of database against applications, techniques and data representations. Extraction of information from image is not rapid with web content mining.

WCM performs any one of following approaches:

- Serial WCM.
- Parallel WCM.

A. Serial Approach

Fig. 3 illustrates design of serial WCM. The processing node would process on link dispatched by URL dispatcher. It was noticed that processing and dispatch time remains constant. For instance consider single link, with B(t) as dispatch time and Q(t) as processing time.

For n link, total time V = n * (B(t) + Q(t)) = n * p = W(n), where p denotes positive constant.

B. Parallel Approach

Fig. 4 illustrates design of parallel WCM. Fig. 5 represents various interconnection network to process nodes, they are hypercube, mesh, star, linear array. Due to topology overhead and communication cost, hypercube interconnection network were chosen as best for processing nodes.

K-cube or K-dimension hypercube is generally used interconnection network in parallel processing and broadly used one. It consist of 2k nodes. 2 nodes were consider as neighbor if their address varies in single bit. K-cube have small diameter which is equal to k. In every step, every node communicate with neighbor node for receiving and passing message which needs log n step to broadcast own message to other nodes.

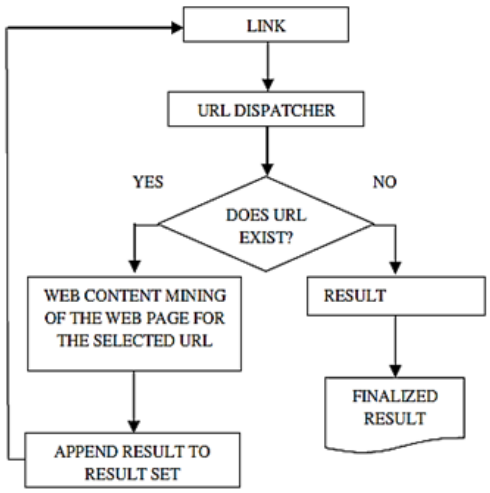

Fig. 3. Serial Approach [21].

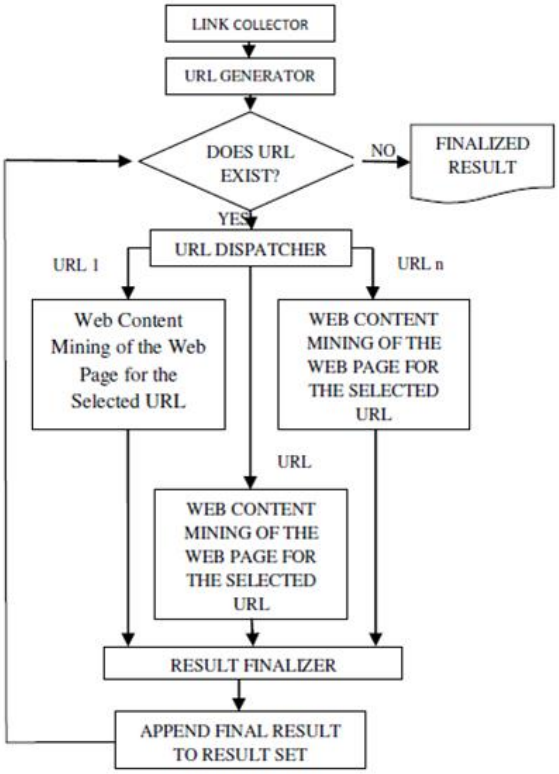

Fig. 4. Parallel Approach [21].

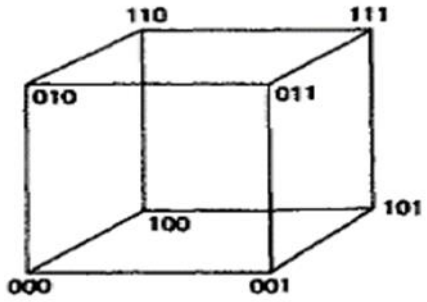

Fig. 5. Representation of Hypercube [21].

V. Various Tools of WCM

In this study, few tools of WCM are considered as most important and reviewed they are:

A. Web Info Extractor

This tool is utilized to mine web data and to retrieve content from web. It retrieve structured or unstructured data from the web page. It doesn't needs complex template rules, where user can browse the webpage and run it [22].

B. Screen-Scraper

It gives (GI) graphical interface facilitating to label URL, data elements to be mined and scripting logic to transverse page and function with mined data. After creating all these items from external language like active server page, PHP, Java, NET, Screen-Scraper could be invoked. This allows scraping of data at regular intervals [22].

C. Mozenda

This tool allow user to mine and control web data. Users set up agents who routinely mine, preserve and publish data to numerous destinations. User can mash-up, recreate and format this tool as per their needs [5].

D. Web Content Extractor

This tool permits users to mine information from several websites like trade directories, economic sites, shopping sites, online public sales and online supplies, etc. Gathered information can transferred to different formats.

E. Automation Anywhere

It is a tool utilized to retrieve web data. Similarly Intelligent automation is a software utilized for information technology and business oriented task.

F. SCRAPY

It is best tool utilized by data professional for learning. It is open source and free software which is written in python language to mine data from websites. It supports information processing and data mining applications to extract structures and crawling of data [23].

G. WEKA

Is a set of machine learning algorithm which consists of clustering, classification, data pre-processing and association rule for data mining applications. The above said algorithm can be used directly with data set. WEKA (Workbench) consist of collation of various tools for visualization and consist of algorithm for data analysis and predictive modelling with GUI (Graphical User Interface) for simple access to their functionality [24, 25].

H. Rapid Miner

It is software environment utilized by same company which provides a combined platform for text mining, data mining and machine learning for business analytics. It is utilized for industrial, application development, rapid prototyping, education training, research, commercial applications and also supports all process in data mining. It utilize client/server module with service provided as cloud infrastructure.

I. ORANGE

It is a machine learning and component dependent data mining software suite, featuring visual programming front-end for explorative visualization and data analysis, libraries and python bindings for scripting. It consists of set of components for pre-processing of data, exploration techniques, modelling, filtering, and feature scoring and model evaluation.

J. KNIME

It is an integration, reporting and open source environment. It's mostly utilized in pharmaceutical research and also in other domain like financial data analysis, business intelligence, and customer data analysis.

K. OCTOPARSE

This tool mimics human behavior to mine data and facilitates users to process website which needs login. IT supports developer to mine all kinds of hyperlinks from

websites. This provide users an easy way to automate 100 of IPs and also provide various improvised options such as built-in XPath tools, Ajax timeout, etc. It crawl data for purpose of web search for particular request and transmit to structured data successfully.

Through this review, various web content mining (WCM) tools as well as its applications are analysed for the extraction of relevant information from the corresponding web page. Web Info Extractor, Screen-Scraper, Mozenda, Web Content Extractor, Automation Anywhere, SCRAPY, WEKA, Rapid Miner, ORANGE, KNIME and OCTOPARSE are the major WCM tools explored in this analysis. Various studies used different techniques of WCM which are Table I comparatively analysed and are explored to be content mining method on the basis of web-text mining, augmented information support etc. Applications of WCM are also revealed through this analysis and it has been found that automatic citation and indexing could be performed with the use of digital library with web mining methods; it can also be utilized to develop, arrange, classify and group most needed information which exist on internet. It also finds its relevance related to user query. On the other hand, challenges associated with WCM are discussed that mainly exhibits long user response time, tedious to execute the queries on non–homogenous data, security remains a major issue in WCM. Most of the analysed studies focused on various terms such as content formation, refusal index, pages number per visit, residence time etc. Quantitative analysis has not been concentrated on many studies. Hence, this study suggests focusing on quantitative analysis and integration of more number of WCM for providing a better perception regarding the significant web patterns.

TABLE I. COMPARATIVE ANALYSIS OF VARIOUS STUDIES ON WEB CONTENT MINING

S.no	Author	Description and methodology	Comments
1.	[26]	The study suggested a web content mining method on the basis of web-text mining to improve the efficiency. Further, the technique of AIS (Augmented information support) was employed to science based websites and developed AIS4XSSC text-mining system. This developed system was analysed for its efficiency, and the major functions are discussed.	The analysed results depicted that the augmented information support technology could efficiently extract the information from huge amount of web texts. Further, the suggested system can effectively improve the information retrieval and could provide valuable information to the users.
2.	[27]	This study suggested as well as implemented a novel web based solution called DAMIS, which was inspired by the cloud. Further, this enables the numerous data mining more effective and simpler for the business intelligence professionals and scientists.	This was beneficial for solving the dimensionality reduction, clustering and classification problems. The suggested solution has broad range of applications as well as provides in-depth information.
3.	[28]	The study suggested a support method as content lifecycle stages in the web systems.	Assists the implementation of web content support. This model described the information resources process in business system followed by the simplification of the automated support technology.
4.	[29]	The study attempted to overcome clustering problem through web document modelling in accordance with the labelled graphs. The study reduced the computational complexity by using the suggested algorithm.	This method promotes the modelling of the web documents without the elimination of contextual data followed by the clustering of these graphical objects with the suggested clustering algorithm.
5.	[18]	This paper provided a comprehensive description of the tools employed in web mining	The study failed to provide description regarding the OCTOPARSE tool. With the help of this study analysis of the stability, retention and usability of various tools could be performed.

VI. Applications

The following section describes few significant applications of web content mining. In general the cloud user utilize the web mining to extract relevant information form the cloud servers. E. commerce sites use these methods to obtain information regarding the specification of the corresponding products. Various search engines enable the user to explore billions of data and assign ranks for all the pages. Based on such ranking and user queries, the search engine orders and publish the pages. Further these techniques would help us to track the individual web sessions more effectively thereby providing valuable information regarding the user behaviour. The personalization plays a vital role for maintaining the confidential and personal information of the user.

Furthermore automatic citation and indexing could be performed with the use of digital library with web mining methods. Electronic services such as e-banking, online knowledge management, blog analysis, social networking, and personalization are used for providing recommendation to the customers. It is used in the crawling of social web platforms such as YouTube, Facebook and Flickr for the verification of the sociological concepts on a large scale for checking the complexity of the mathematical model. Apart from that terrorism is regulated through web mining since the terrorists utilize blogs, social networking, website forums, and virtual world for the exchange of information. By analyzing carefully, the data enable to regulate terrorism.

In addition, WCM is utilized to develop, arrange, classify and group most needed information which exist on internet. It also find its relevance related to user query. Most useful for online marketing by improved investigation of information on web. Users refine information with use of WCM, track online behavior of user, examine website productiveness and supports digital marketing over intelligence of product price, etc. Users having similar interest are club together and depending on examining the content they were posted on social media sites. Web mining technique is utilized to optimize online content since large quantity of content is added to WWW (World Wide Web) every day. Cloud managing with multiple videos, images and files must need to be optimized [30].

A. Limitations and Challenges

- When investigating the limitations of web content mining approaches, the following points are observed.
- Very long user response time. Accumulation of massive web resources leading to traffic.
- Any kind of solution for improvising the efficiency and quality might lead to the rise in bandwidth and cost.
- In case of web catching system, improper updation might cause bottleneck in servers, stale data and increase in the number of users.
- Still there exists no standardized tool to deal with over fitting, under fitting and oversampling of the data.
- There exists severe complexity in the elimination of duplicate data.
- The web content mining for time series data followed by its scaling up procedure seems to be complex.
- The mining of hyperlink network and structures are found to be difficult.
- It is tedious to execute the queries on non – homogenous data.
- Overall security remains a major issue.

In addition to that, mining one server is not useful so it needs numerous counts of servers to process and mine useful information. Special hardware and software were obligatory to extract terabytes of datasets. There is also chance of deleting new data from web with use of automatic cleaning process. There will be restricted inquiry interface, scope and customization to every clients. Some user needs more data than others or sometimes less than others or sometime there is need for extended data refining than prescribed. In some case its tedious task to search important information on web since it changes dynamically.

VII. Conclusion and Future Work

This paper deliberates the web content mining concepts, its classifications, and the associated major techniques. This survey also describe the data mining algorithm and methods for discovering the useful information from web. The application, recent trends followed by the limitations and challenges are also discussed in this paper. This paper also provide insight about the conventional web mining algorithms in simplifying the processing methods. This research explores various techniques, approaches and tools of WCM along with its applications and issues. Here researcher also establishes few objective criteria for comparison of WCM tools. WCM scans webpage content like text, video, audio, image and HTML content. Its output is supplied to search engine to develop most relevant knowledge. Since webpage consist of dynamic data it is a tedious process to extract data which is updated daily. In future, WCM must enhance its technique, approaches and tools to enhance scalability, usability and users retention. The future work focused on the integration of more number of web content mining for providing a better perception regarding the significant web patterns.

References

[1] N. Pradhan and V. Dhaka, "Comparison-based study of pagerank algorithm using web structure mining and web content mining," in Smart Systems and IoT: Innovations in Computing, ed: Springer, 2020, pp. 719-729.

[2] I. M. Shahid and A. K. Srivastava, "An Integrated approach for process in knowledge discovery on World Wide Web (Internet) using web mining."

[3] S. M. Huded, S. Balutagi, and A. Ranjan, "Mapping of literature on data mining by j-gate database," 2019.

[4] S. Taha Ahmed, R. Al-hamdani, and M. Crook, "Studying of Educational Data Mining Techniques," International Journal of Advanced Research in Science, Engineering and Technology, vol. 5, pp. 5742-5750, 2018.

[5] E. T. John, B. Skaria, and P. Shajan, "An Overview of Web Content Mining Tools," Bonfring International Journal of Data Mining, vol. 6, pp. 01-03, 2016.

[6] T. A. Al-asadi, A. J. Obaid, R. Hidayat, and A. A. Ramli, "A survey on web mining techniques and applications," International Journal on

Advanced Science Engineering and Information Technology, vol. 7, pp. 1178-1184, 2017.

[7] S. Saini and H. M. Pandey, "Review on web content mining techniques," International Journal of Computer Applications, vol. 118, 2015.

[8] S. Vijiyarani and E. Suganya, "Research issues in web mining," International Journal of Computer-Aided Technologies (IJCAx), vol. 2, p. 55, 2015.

[9] S. N. Kumar, "World towards advance web mining: A review," American Journal of Systems and Software, vol. 3, pp. 44-61, 2015.

[10] A. Kumar and R. K. Singh, "A Study on Web Content Mining," IJECS, vol. 6, pp. 20003-20006, 2017.

[11] K. Srinath, "An Overview of Web Content Mining Techniques," 2017.

[12] X. L. Mary, G. Silambarasan, and M. phil Scholar, "Web content mining: tool, technique & concepts," Int. J. Eng. Sci, vol. 7, p. 11656, 2017.

[13] M. J. H. Mughal, "Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview," Information Retrieval, vol. 9, 2018.

[14] S. Shanthi, "Survey on web usage mining using association rule mining," International Journal of Innovative Computer Science & Engineering, vol. 4, 2017.

[15] S. Jayaprakash and D. Owusu, "Survey on Web Usage Mining using Association Rule Mining."

[16] N. Parmar, V. Richhariya, and J. P. Maurya, "An Exploratory Review of Web Content Mining Techniques and Methods," International Journal of Advanced Research in Computer and Communication Engineering, vol. 5, pp. 144-148, 2016.

[17] V. M. Alexandrino, G. Comarela, A. S. da Silva, and J. Lisboa-Filho, "A Focused Crawler for Web Feature Service and Web Map Service Discovering," in International Symposium on Web and Wireless Geographical Information Systems, 2020, pp. 111-124.

[18] R. H. Salman, M. Zaki, and N. A. Shiltag, "A Studying of Web Content Mining Tools," Al-Qadisiyah Journal Of Pure Science, vol. 25, pp. 1-16, 2020.

[19] S. Vijayarani and A. Sakila, "Multimedia mining research-an overview," International Journal of Computer Graphics & Animation, vol. 5, p. 69, 2015.

[20] C. C. Aggarwal, "Mining text data," in Data Mining, 2015, pp. 429-455.B. Panda, S. N. Tripathy, N. Sethi, and O. P. Samantray, "A comparative study on serial and parallel web content mining," International Journal of Advanced Networking and Applications, vol. 7, p. 2882, 2016.

[21] V. Bharanipriya and V. K. Prasad, "Web content mining tools: a comparative study," International Journal of Information Technology and Knowledge Management, vol. 4, pp. 211-215, 2011.

[22] D. Ahamad, D. Mahmoud, and M. Akhtar, "Strategy and implementation of web mining tools," International Journal of Innovative Research in Advanced Engineering, vol. 4, pp. 01-07, 2017.

[23] N. Sharma and K. Bansal, "Comparative study of data mining tools," Journal of Advanced Database Management & Systems, vol. 2, pp. 35-41, 2015.

[24] K. Rangra and K. Bansal, "Comparative study of data mining tools," International journal of advanced research in computer science and software engineering, vol. 4, pp. 216-223, 2014.

[25] C. Li, "Research on an Enhanced Web Information Processing Technology based on AIS Text Mining," Recent Advances in Electrical & Electronic Engineering (Formerly Recent Patents on Electrical & Electronic Engineering), vol. 14, pp. 29-36, 2021.

[26] V. Medvedev, O. Kurasova, J. Bernatavičienė, P. Treigys, V. Marcinkevičius, and G. Dzemyda, "A new web-based solution for modelling data mining processes," Simulation Modelling Practice and Theory, vol. 76, pp. 34-46, 2017.

[27] V. Vysotska, V. B. Fernandes, and M. Emmerich, "Web Content Support Method in Electronic Business Systems," in COLINS, 2018, pp. 20-41.

[28] K. Phukon, "Incorporation of contextual information through Graph Modeling in Web content mining," Indian Journal of Science and Technology, vol. 13, pp. 4573-4578, 2020.

[29] N. Satish, "A Study on Applications, Approaches and Issues of Web Content Mining," International Journal of Trend in Research and Development, vol. 4, pp. 41-43, 2017.