A Systematic Review of Current Trends in Web Content Mining

Makinde Opeyemi Samuel, Afolabi Ibukun Tolulope, and Oladipupo Olufunke Oyejoke
Department of Computer Science, Covenant University

ibukun.fatudimu@covenantuniversity.edu.ng

Abstract. Knowledge in web documents, Relevance ranking of webpages and so on are some of the under-researched areas in web content mining (WCM). Apart from the general data mining tools used for knowledge discovery in web, there have been few attempts at reviewing WCM and these were from the perspective of the methods used and the problems solved but not in sufficient depth. This existing literature review attempts does not also reveal which problems have been under-researched and which application area has the most attention when it gets to WCM. The goal of this systematic review is to make available a comprehensive and semi-structured overview of WCM methods, problems and solutions proffered. To provide a comprehensive literature review on this subject, 57 publications which include journals, conferences proceeding, and workshops were considered between the periods of 1999-2018. The findings reveal that updating dynamic content, efficient content extraction, eliminating noise blocks etc remain the most prominent challenges associated with WCM with a very high attention on solving these problems in a more efficient manner. Also, most of the solutions proffered to the problems still come with their various limitations which make this area of research fertile for future research. Caching dynamic web data. With regard to content, the techniques used for content extraction in WCM consist of used Data Update Propagation (DUP), Association rule, Object Dependence Graphs, classification techniques, Document Object Model, Vision-Based Segmentation, Hyperlink-Induced Topic Search and so on. Finally, the study revealed that WCM has been mostly applied to general websites which include random webpages seeking to extract specific parameters. The review was able to identify the limitations of the current research on the subject matter and identify future research opportunities in WCM.

1. INTRODUCTION

In the last decades, the web has become the largest publicly accessible data source in the world because of its swift growth. The web possesses some distinctive characteristics which make extracting interesting information and knowledge from it a captivating and challenging task. Some of these characteristics are; huge amount of information, data of different types, heterogeneous information, linked information, noisy information, dynamic information, the web is about business, commerce and a virtual society. These unique characteristics gave an opportunity for data mining and machine learning techniques for extracting useful information from the web[1].

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd
Therefore, extracting useful information and patterns from Web hyperlink structure, page content, and usage data using data mining and machine learning techniques is known as web data mining. Some of the useful data mining and machine learning techniques for web mining are classification analysis, association rule learning, clustering analysis, support vector machine algorithm, artificial neural networks[2]–[5]. Web mining task can be categorized into three: web structure mining, web usage mining and WCM [1], [6]–[9].

Web structure mining involves discovering useful information or knowledge from hyperlinks. For instance, analysing the nodes and connection structure of a website using graph theory [5], [10]. Two things are discovered: the website structure which shows its interconnectivity to other websites and the document structure of the website specifically, relating to the interconnectivity of pages within the website[11]. Web structure mining minimizes digressive search results and the inability to index the massive amount of data available on the World Wide Web (WWW).

Web usage mining entails extracting patterns and data from server logs to understand client or user actions including where they are from, the number of clients or users who clicked an item on the web page and the sort of activities being done on the website [12]. Large volumes of data are generated and collected daily by organizations. Majority of the data are automatically generated by web servers and collected in logs, which can be analysed to determine the effect of promotion, cross marketing strategies and so on. Also, helping an organization to better structure a website to increase its presence [5].

WCM involves mining useful information from Web pages contents. The contents of webpages are usually textual, images and audio or video data[13], [14]. WCM could be achieved by cluster and classify web pages automatically based on their topics (postings of forums, product descriptions, specific words) to unveil hidden patterns and knowledge. For instance, customers opinions can be unveiled by product reviews and postings made by them on the web.

In this study, the overarching focus of the systematic review is on WCM and seven main research questions are established systematically to evaluate the trends and direct future research areas in WCM. Several journal articles and conference papers on WCM were reviewed to establish the discovered trends.

2. WEB CONTENT MINING

WCM can be broken down into search result and webpage content mining[6]. In web page content mining useful hidden patterns are extracted from web pages contents, while search result mining involves using the contents of web pages to rank web pages for a search query.

In carrying out WCM task, there are challenging issues in automated discovery processes which are due to the varied structure in web resources [15]. Hence, the extraction approaches depend on the web page content data type. The web page contents can be categorized into unstructured data type, which entails unstructured text mining methods such as clustering, summarization, information extraction, information visualization and topic tracking; Structure data type, which entails structure data mining techniques such as page content mining, web crawler and wrapper generation; Semi-structured data, which entails semi-structure data mining techniques such as web data extraction language, object exchange model and top down extraction; and lastly multimedia data type which entails multimedia data mining methods such as shot boundary detection, multimedia miner, SKICAT and colour histogram matching[8].

Useful information such as forum postings, the product description is extracted from web page content as a result of mining processes for several reasons in meeting specific needs. For instance, e-commerce websites, blogs and forums can be mined to discover consumer opinions.

Application areas such as recommender systems, internet search engines, business intelligence, uncovering web robots etc. involve the use of different WCM techniques.
Challenging issues in automated discovery processes are due to the unavailability of structure in web resources [15], hence, various techniques are used for different data types – unstructured data, which is not organized and does not have a specific data type; structured data which is organized and is managed by a Database Management System (DBMS) and semi-structured data are partly organized.

A typical WCM architecture comprises of various processes. These processes vary in relation to the expected result. According to [16], a typical WCM architecture consists of six main components which includes; user, webserver, web content, data mining techniques, recommendation engine and recommendations. The user searches a web server through a query or by maneuvering links around the web pages; the web contents express the users’ interest. Specific contents frequently reviewed by the user are noted and organized in clusters or discovered trends or pattern, using certain data mining algorithms, such as hierarchical agglomerative clustering, semantic characterization and keyword extraction and so on.

3. RESEARCH QUESTIONS

The aim of this research is to answer the following research questions (RQs):

RQ1: What WCM application domain has been explored in literature?

- The motivation for this question is to identify the application domains that have benefited from WCM and why.

RQ2: What kind of methods are the most used for WCM?

- The motivation for this question is to discover trends applied methodologies in web mining thereby establish the state of the art methodologies.

RQ3: What is the dominant WCM problem?

- The motivation for this question is to establish the open problems in WCM.

RQ4: What solutions were offered to the problems identified?

- The motivation of this question is to discover existing solution to the open problems.

RQ5: Which WCM problem has been under-researched?

- The purpose of the answer to RQ5 is to identify gaps for future research.

RQ6: Which journal is dominant in the field of WCM?

- The main motivation for this is to discover most important WCM journal.

RQ7: What is the percentage of publications published after the year 2013?

- This question is directed at discovering the recent contribution trend in WCM.

4. METHODOLOGY

The research questions are structured to express the content of literature review particularly following the approach of [17] and [18]. Scopus database, IEEEExplore and CiteSeerx were the main source for the publication due to their richness in content as regards WCM publications. The search in the earlier mentioned databases was on WCM in the article title, abstract or keywords. Subsequently, the search string was modified to include “web content extraction” and “content extraction”. 80% of the papers used were found in the Scopus database.

To maximize the Scopus database in retrieving the publications we used, we combined keywords to test synonyms used in literature to uncover the variety of publications in WCM. The following restriction help to define the boundary of our study; (i) source type (i.e. journal and conference papers, (ii) publication year - before and including the last quarter of 2017, and (iii) Scopus’ subject area of Computer Science.
4.1. Inclusion criteria:
(1) Web data, WCM, problems, typical application and future trends were guides to arriving at the search criteria and the major topics of the publications, (2) Only the most recent publications were selected in cases of multiple publications.

4.2. Exclusion criteria:
(1) General WCM procedure publications were excluded because they did not focus on solving a problem as regards to WCM. (2) Publications that focused on general WCM tools we also excluded. (3) Journal articles that were not accessible online are excluded. (4) Publications on web mining that focused on solving web structure mining and web usage mining were not used.

Relevant criteria were applied to the titles and abstract of the papers, and in cases where it was not clear enough, the whole publication was scanned to determine the inclusion and exclusion papers for this review. The outcome resulted in 57 publications included in the next research steps. These 57 publications were selected from a total of 78 which was retrieved before applying the inclusion criteria. The year of the publications selected ranged from 1998 to 2018.

5. RESULT AND DISCUSSION

RQ 1: What WCM application area has been treated in literature? To answer the first research question, the application areas are divided into General websites – i.e. random webpages seeking to extract specific parameters, Sports – includes papers that are sport related e.g. Olympics, etc., publishing – content extraction for publishing reasons, news – papers showing content extraction from news webpages, Search Engines – content extraction based on the search results of search engines, multi-language websites – paper that show content extraction from non-English webpages i.e. Chinese, Arabic and Vietnamese webpages, Educational – online course extraction for recommendation, Entertainment – music and video extraction, Social media – extracting twits for various purposes, Tourism – extracting tourist data from tourism related webpages.

From the diagram in Figure 1, it is obvious that general website data has benefited extensively from WCM with a percentage of 53% followed by News with 21%. This result is supported with these references [19]–[26].
RQ2: What kind of methods are the most used for WCM?
To answer this question, we focus on reviewing the methods employed for the most important phase of WCM which is web content extraction. This is because the result of this phase determines the result of the knowledge discovery phase.

The methods used according to literature are explained as follows; [27] used Data Update Propagation (DUP) for web content extraction, [28] used Association Rule for web content extraction but did not cover many Hoover websites. [29] used Object Dependence Graphs (ODG) for efficient web content extraction. [21], [25], [30]–[35] used the classification technique for effectual content extraction, issues such as scalability, noise etc. were discovered at various points. [36]–[38] used Natural Language Processing (NLP) to efficiently extract contents from different webpages. [23]–[25], [39]–[53] used Document Object Model (DOM) to optimize web content extraction. [19], [22], [54] used Vision Based Segmentation for efficient content extraction. [55] used the Hyperlink Induced Topic Search (HITS) to efficiently extract web contents. [20], [30], [54], [56]–[60] showed that clustering can be used for effective web content extraction. [61] used Page Content Rank for content extraction. [62]–[66] used various extraction algorithms, however, noise was still found after they were used. [67] used Key Information Method for web content extraction. [22] used Recognizing Information Page Blocks (RIPB) for effective content extraction. [68], [69] used content code (Content Code Vector (CCV) and Content Code Blurring (CCB)) to extract contents from webpages. [70]–[72] used different web crawling algorithms for content extraction. [71] used Categorization and Information visualization for efficient web content extraction. [30], [73] showed that hashing can be used for effective content mining. [32] used shallow similarity analysis to extract web contents. [73] used block splitting to effectively mine contents from webpages. [26] used Sentiment and Keyword Analysis to efficient extract contents from webpages. [74] used RSS Feed for web content extraction. [75] used a Video text detection-based algorithm to mine contents from webpages. [76] used Name Entity Recognition (NER) to mine twit location from twitter post. From the review of the methods employed in WCM it is obvious that DOM (Document Object Modeling) is mostly used to optimize web content extraction in
RQ3: What is the dominant WCM problem?

The specific pitfalls observed with respect to various WCM researches include the following; data cleaning in WCM is very challenging due to the availability of additional noise sources from the inadequacies of feature extractors[28].

Inability to separate webpages that comprise of tags like DIV other than than the table tag and only Chinese news webpages were considered for experimentation[46]. [55]used Hyperlink Induced Topic Search (HITS) algorithm which works by finding hub and authority page, but it becomes difficult for it to find out those hub web pages that have few authority pages linked with it.

Lihui and Chue in [20] introduced a technique called Semantic Visual Document (SVD) for representing web documents. However, the problem with this approach is that the number of words for the query expansion was selected via intuition and no research was carried out to test their validity.

Smizansky et al in [61] proposed a method called Page Content Rank (PCR), it combines several heuristics that are important for analysing the content of web pages. Nevertheless, the time complexity of obtaining the starting set of pages is a weak point of the PCR implementation. In [49], Pan et al. suggested a technique for web page content extraction based on statistics and link density, yet, the benchmark values in this paper don't generally adjust to some news pages, for example, stock news, so it is still hard to locate an arrangement of all-inclusive parameters. Weninger et al. in [60] developed a method that uses tag ratios. To compute the ratio of several non-HTML tag characters to the total number of HTML tags available, tag ratios are used. However, the limitation of this method is the indentation of the codes in webpages which can provide different tag ratio values. In [39] Bhardwaj and Mangat presented an improved word to leaf ratio (WLR) approach was proposed, but the algorithm does not extract content from the previously mentioned tags if they are used in the background of <div> tag’s property. Mathiyalagan in [32] proposed a framework aimed to enhance the existing string-matching plagiarism detection approach with similarity analysis techniques. In this study the program could not detect similar repeating sections of the same document.

Koltringer and Dickinger in [26] presented an automated WCM approach. The accuracy of the categorization approach is limited due to the human component. The future challenge is to identify processes to build measures and models to help understand how social media, news media and other influences impact destination brand image.

In [59] Menaka and Nagadeepa presented a method that used clustering to extract images from web pages was proposed. However, the training data set was small compare to mostly used pictures in literature. Detecting and updating dynamic content – dynamic contents are produced by most websites ranging from ecommerce sites to sport sites. Creating and updating data on these websites are problems for efficient web content extraction [29], [37]. Efficient content extraction – due to the immense increase of web data, research to explore different fields becomes important, many of these require adequate content for optimizing productivity and efficiency[57], [60], [75]. Detecting and extracting web content structure – webpages contain a lot of contents based on the topic of the webpages which vary from navigation to contact information, also, a webpage may contain many irrelevant topics, these reduces the performance of the efficient content extraction [19], [44]. Identifying and extracting web content blocks – unnecessary blocks are contained in webpages, they include adverts, unnecessary links. Removing these blocks and extracting important informative content blocks becomes necessary[45], [62]. Eliminating noise blocks – noise blocks contain unwanted contents (adverts etc.), they affected content extraction negatively[21], [69], [73]. Others include the following: caching dynamic web data [27], representing knowledge in web documents [20], relevance ranking of webpages [61], enhancing string-matching plagiarism detection approach [32], web crawling and running Natural Language Processing (NLP) parallelly[38], classifying posts as content related or not [35], web content credibility [58]
Figure 2 reveals that the dominant research problem is to achieve an efficient extraction of web contents. In as much as most researchers focused on solving this problem, it still possesses a major problem to researchers as identified in RQ4.

RQ4: What solutions were offered to them and direction for future research?

Some of the proposed solutions to the above problems is described as follows; in the process of information extraction, information from wrapped websites as a source of training data to improve their extractors can be a viable option for further research [28]. [37] proposed adopting the crawler for locating news in Spanish; improve the document representations by constructing logic predicates or conceptual graphs as formal text representations; extend the generalization methods in order to discover other kind of patterns such as: associations, clusters and deviations; specialize the whole process to some domain such as economy or politics. [19] proposed a new approach of extracting web content structure based on visual representation. The use of this plan for versatile content conveyance to encourage better web browsing on cell phones can be considered for future works. [62] suggests that considerations should be on the algorithms as part of a system that crawls web pages, and primary content blocks are extracted from it. It is also imperative to look at the primary content and identify heuristic algorithms to identify the semantics of the content to generate mark-up. In [21], Li and Ezeife proposed a framework, WebPageCleaner, for wiping out noise blocks from web pages for enhancing the precision and productivity of WCM. In [67], Wang, Lu and Zhang suggested incorporating the KIM (Key Information method method) into classification or integration model building to reduce the cost of labelling large training samples is a necessary focus for researchers. [68] suggests that for further work on Content Code Blurring (CCB), researchers will have to include fine tuning of the parameters, combining it with other methods to achieve better results and to incorporate some notion of the Document Object Model (DOM) block elements to enhance recognition of structurally related parts of a document. [49] recommends that further investigation concerning the advancement of search engine interfaces which incorporate concept recovery and extension of semantic and equivalent words are imperative. [23] believes that SDOM can be expanded to a more complicated semantic
environment. More research into integrating semantic into DOM tree is necessary. [60] stated that there is a need for further investigation on the clustering algorithm used in Content Extraction via Tag Ratios (CETR). [71] suggested that it is imperative to compare focused crawling and direct search using a search engine. [77] recommends that for further research, more focus should be made to other algorithms of clustering to improve performance. [72] suggests that more focus should be on refinement of the requirements for DDNA used as a tool in the process of dynamic web data mining and on the problem of interaction between web sites and experiential knowledge base repositories developed during crawling experiences. They also recommended research into the possible application of proposed combination of DDNA and web crawlers as smart decision support assistant in data mining processes. [43] suggest that by collecting the concentration of links in cases where the links contain one or two words, it could be possible to find the menus of the webpage. [25], [39] for further study, suggested adopting their approach for a crawler to efficiently extract informative content from the web, also, it is imperative to automatically infer different data, including name, price, comments, and technical product details, from shopping web domains. [32] suggests that making their program detect similar repeating sections in the same document could be researched on. [40] recommends that researchers could consider incorporating information to their method could be vital and on event information generation. [66] aim to use the extracted content from the BiLEx algorithm as the input source of data to a classification system already built, then improve the system to create a simple search engine. [26] suggests that further conceptualization about how to employ content mining for market research is necessary, and a longitudinal study monitoring online destination representation is an interesting future research avenue. In [57], Huang et al proposed a clustering-based web content extraction algorithm. The algorithm can be adopted into various WCM applications. [70] suggest that for further studies, their proposed procedures can be followed to develop Dynamic Relevance Search Engines (DRSE) that combines personal, social and local data to show relevant results. [35] suggest that their model could be used to group starting posts and replies to yield different analytics about the relatedness of the contents of threads.

According to literature, WCM is a blossoming research area with many opportunities. The above-mentioned suggestions show that the techniques, algorithms and methods presented can be utilized in various ways for implementing ground-breaking research.

RQ 5: Which WCM problem has been under researched?

| S/N | Problems                                              | Paper Count |
|-----|-------------------------------------------------------|-------------|
| 1.  | Detecting and updating dynamic web content            | 2           |
| 2.  | Efficient content extraction                          | 31          |
| 3.  | Detecting and extracting web content structure        | 5           |
| 4.  | Identifying and extracting content blocks             | 5           |
| 5.  | Eliminating Noise blocks                              | 7           |
| 6.  | Others Caching dynamic web data                      | 1           |
|     | Representing knowledge in web document                | 1           |
Based on the review conducted, a lot of WCM problems have been under researched, therefore enabling the domain a practical choice for research. According to this study, the following WCM problems were under researched:

- Caching dynamic web data [27]
- Representing knowledge in web documents [20]
- Relevance ranking of webpages [61]
- Enhancing string-matching plagiarism detection approach [32]
- Web crawling and running Natural Language Processing (NLP) parallelly [38]
- Classifying posts as content related or not [35]
- Web content credibility [58]

RQ6: Which journal is dominant in the field of WCM?

Table 1 gives the summary of the frequency of publications in specific journal, revealing how often they publish content mining research. All journal/conference or workshops that have only one publication have been excluded from the table.

Table 1 shows that in WCM, the journal of Information Processing and Management was dominant across the papers that were reviewed for this study, followed by the IEEE Transactions on knowledge and Data Engineering journal.

| Journal/Conference/Workshop                              | Count |
|----------------------------------------------------------|-------|
| IEEE Transactions on Knowledge and Data Engineering      | 3     |
| IEEE Conference on Computer Communications               | 2     |
| ACM SIGKDD International Conference on Knowledge Discovery and Data Mining | 2     |
| Asia-Pacific web conference                              | 2     |
| Information Processing and Management                    | 5     |
| International Workshop on Database and Expert Systems Applications | 2     |
| Expert Systems with Applications                         | 2     |
RQ7: What is the percentage of publications published after year 2010?

Figure 3 reveals that 36% of the total publications used belong to the 2013 till date category, meaning that there is increase in research on WCM in recent years.

6. CONCLUSION AND FUTURE WORK

In this study, a comprehensive systematic review and analysis of paper published in journals, conference proceedings and workshops in WCM for 20 years was conducted. The aim was to assess the current research articles in WCM to discover the trend in contributions and the open problems in WCM. Through the review the following were observed:

- Application of WCM to general websites has had the most attention followed closely by websites contains one kind of news or the other (Debnath et al., 2005; Kakol et al., 2017; Li & Ezeife, 2006; Sambanthan, & Dhenakaran, 2017).
- Since the knowledge discovery techniques in WCM remain the same as data mining, the major uniqueness’ of methods used for WCM is in the pre-processing of web content, particularly content extraction. The techniques commonly used for content extraction in WCM are Data Update Propagation (DUP), Association rule, Object Dependence Graphs, classification techniques, Document Object Model, Vision Based Segmentation, Hyperlink Induced Topic Search and so on [33], [47][78].
- Detecting and updating dynamic content, efficient content extraction, eliminating noise blocks etc are some of the problems associated with WCM[40]. However, there are more contribution to providing a more efficient content extraction in literature.
- Most of the solutions proffered in literature to the discovered WCM problems still come with their various limitation, therefore, there is need to keep looking for better solutions [35]. [57][79].
- Caching dynamic web data, representing knowledge in web documents, Relevance ranking of webpages, enhancing string-matching plagiarism detection approach were identified to be under researched, thus gives opportunity for future research in the areas.

The above observations imply a very rich research opportunity in WCM. It reveals the limitations of current researches in WCM studies and identifies future research opportunities.

References
[1] B. Liu, Web Data Mining. 2011.
[2] C. Dinuca, “Using web mining in e-commerce applications,” Ann. Ser., 2011.
[3] P. Menaka and A. Prathamadevi, “A Survey on Web Mining and Its Techniques,” Int. J. Adv.
Res. Comput. Sci. Softw. Eng., vol. 5, no. 9, pp. 276–280, 2015.

[4] T. M. Modi, B. Patel, and P. Panchal, “A Survey: Web Recommendation Using Different Approaches of Web Mining Techniques.,” Int. J. Adv. Found. Res. Comput., vol. 2, no. January, pp. 11–19, 2015.

[5] M. Yaduv and P. Mittal, “Web Mining: An Introduction,” Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 3, no. 3, pp. 683–687, 2013.

[6] S. Deepa and S. Harirahan, “Web Content Mining: Issues and Challenges,” IUP J. Inf. Technol., vol. 8, no. 3, pp. 64–72, 2012.

[7] P. Desikan et al., “Web Mining for Business Computing,” Handb. Inf. Syst., pp. 1–36, 2009.

[8] F. Johnson and S. Kumar Gupta, “Web Content Mining Techniques: A Survey,” Int. J. Comput. Appl., vol. 47, no. 11, pp. 44–50, 2012.

[9] N. V. Nair, “Mining the Web for Managing Web Content,” IUP J. Syst. Manag., vol. 7, no. 3, pp. 7–16, 2009.

[10] J. Srivastava, P. Desikan, and V. Kumar, “Web Mining — Concepts, Applications, and Research,” Int. Res. J. Comput. Sci., vol. 3, no. 1, pp. 8–13, 2016.

[11] P. Chopra and M. Ataullah, “A Survey on Improving the Efficiency of Different Web Structure Mining Algorithms,” Int. J. Eng. Adv. Technol., vol. 2, no. 3, pp. 296–298, 2013.

[12] Y. H. Tao, T. P. Hong, and Y. M. Su, “Web usage mining with intentional browsing data,” Expert Syst. Appl., vol. 34, no. 3, pp. 1893–1904, 2008.

[13] D. Yang and J. Song, “Web content information extraction approach based on removing noise and content-features,” Proc. - 2010 Int. Conf. Web Inf. Syst. Mining, WISM 2010, vol. 1, pp. 246–249, 2010.

[14] S. Kundu, “An intelligent approach of web data mining,” Int. J. Comput. Sci. Eng., vol. 4, no. 5, pp. 919–928, 2012.

[15] B. S. Anami, R. S. Wadawadagi, and V. B. Pagi, “Machine Learning Techniques in Web Content Mining: A Comparative Analysis,” J. Inf. Knowl. Manag., vol. 13, no. 1, p. 1450005, 2014.

[16] C. Joshi and U. Kumar, “A Novel Approach towards Integration of Semantic Web Mining with Link Analysis to Improve the Effectiveness of the Personalized Web,” Int. J. Comput. Appl., vol. 128, no. 11, pp. 1–5, 2015.

[17] B. Kitchenham, “Procedures for Performing Systematic Reviews,” Jt. Tech. Rep. Softw. Eng. Group, Keele Univ. United Kingdom Empir. Softw. Eng. Natl. ICT Aust. Ltd, Aust., 2004.

[18] J. Webster and R. T. Watson, “Analyzing the Past to Prepare for the Future: Writing a Literature Review,” MIS Q., vol. 26, no. 2, pp. xi–xxiii, 2002.

[19] D. Cai, S. Yu, J. Wen, and W. Ma, “Extracting Content Structure for Web Pages Based on Visual Representation,” Proc. 5th Asia-Pacific Web Conf. - APWeb ’03, pp. 406–417, 2003.

[20] C. Lihui and W. L. Chue, “Using Web structure and summarisation techniques for Web content mining,” Inf. Process. Manag., vol. 41, no. 5, pp. 1225–1242, 2005.

[21] J. Li and C. I. Ezeife, “Cleaning Web Pages for Effective Web Content Mining,” Dexta 2006, Lncs 4080, pp. 560–571, 2006.

[22] J. Kang and J. Choi, “Detecting informative Web page blocks for efficient information extraction using visual block segmentation,” Proc. - 2007 Int. Symp. Inf. Technol. Converg. ISITC 2007, pp. 306–310, 2007.

[23] F. Yulian, L. Zongwei, X. Yun, and Z. Winston, “A semantic DOM approach for webpage information extraction,” Proc. - Int. Conf. Manag. Serv. Sci. MASS 2009, 2009.

[24] F. Sun, D. Song, and L. Liao, “DOM based content extraction via text density,” Proc. 34th Int. ACM SIGIR Conf. Res. Dev. Inf. - SIGIR ’11, vol. 1, p. 245, 2011.

[25] E. Uzun, H. V. Agun, and T. Yerlikaya, “A hybrid approach for extracting informative content from web pages,” Inf. Process. Manag., vol. 49, no. 4, pp. 928–944, 2013.

[26] C. Költringer and A. Dickinger, “Analyzing destination branding and image from online sources: A web content mining approach,” J. Bus. Res., vol. 68, no. 9, pp. 1836–1843, 2015.
[27] J. Challenger, A Iyengar, and P. Dantzig, “A scalable system for consistently caching dynamic Web data,” *IEEE INFOCOM 99 Conf. Comput. Commun. Proc. Eighteenth Annu. Jt. Conf. IEEE Comput. Commun. Soc. Futur. is Now Cat No99CH36320*, vol. 1, no. e, pp. 294–303 vol.1, 1999.

[28] R. Ghani, R. Jones, and D. Mladenic, “Data mining on symbolic knowledge extracted from the web,” *Data Min.*, no. August, pp. 29–36, 2000.

[29] J. Challenger, A. Iyengar, K. Witting, C. Ferstat, and P. Reed, “A publishing system for efficiently creating dynamic Web content,” *Proc. IEEE INFOCOM 2000. Conf. Comput. Commun. Ninet. Annu. Jt. Conf. IEEE Comput. Commun. Soc. (Cat. No.00CH37064)*, vol. 2, no. May, pp. 844–853, 2000.

[30] P. Agarwal, R. Vaithiyanathan, S. Sharma, and G. Shroff, “Catching the Long-Tail: Extracting Local News Events from Twitter,” *Sixth Int. AAAI Conf. Weblogs Soc. Media*, pp. 379–382, 2012.

[31] C. Bucur, “Using Opinion Mining Techniques in Tourism,” *Procedia Econ. Financ.*, vol. 23, no. October 2014, pp. 1666–1673, 2015.

[32] M. Mathiyalagan, “Mining the Web for Plagiarized Web Content using Similarity Measure Technique,” *Int. J. Res. Sci. Eng. Technol.*, vol. I, no. I, pp. 32–38, 2014.

[33] K. B. Prakash and A. Rajaraman, “Mining of Bilingual Indian Web Documents,” *Procedia Comput. Sci.*, vol. 89, pp. 514–520, 2016.

[34] A. F. R. Rahman, H. Alam, and R. Hartono, “Content extraction from HTML documents,” *Proc. First Int. Work. Web Doc. Anal. (WDA 2001)*, no. July, pp. 7–10, 2001.

[35] A. F. Wise, Y. Cui, W. Q. Jin, and J. Vytasek, “Mining for gold: Identifying content-related MOOC discussion threads across domains through linguistic modeling,” *Internet High. Educ.*, vol. 32, pp. 11–28, 2017.

[36] A. F. R. Rahman, H. Alam, and R. Hartono, “Content extraction from HTML documents,” *Proc. First Int. Work. Web Doc. Anal. (WDA 2001)*, no. July, pp. 7–10, 2001.

[37] A. Mendez-Torreblanca, M. Montes-y-Gómez, and A. López-López, “A Trend Discovery System for Dynamic Web Content Mining,” *Int. Conf. Comput.*, no. 1, p., 2002.

[38] P. Nesi, G. Pantaleo, and G. Sanesi, “A hadoop based platform for natural language processing of web pages and documents,” *J. Vis. Lang. Comput.*, vol. 31, pp. 130–138, 2015.

[39] A. Bhardwaj and V. Mangat, “A novel approach for content extraction from web pages,” *2014 Recent Adv. Eng. Comput. Sci.*, pp. 1–4, 2014.

[40] A. Bhardwaj and V. Mangat, “An improvised algorithm for relevant content extraction from web pages,” *J. Emerg. Technol. Web Intell.*, vol. 6, no. 2, pp. 226–230, 2014.

[41] Y. Biletskiy, J. A. Brown, and G. Ranganathan, “Information extraction from syllabi for academic e-Advising,” *Expert Syst. Appl.*, vol. 36, no. 3 PART 1, pp. 4508–4516, 2009.

[42] Y. Guo, H. Tang, L. Song, Y. Wang, and G. Ding, “ECON: An approach to extract content from web news page,” *Adv. Web Technol. Appl. - Proc. 12th Asia-Pacific Web Conf. APWeb 2010*, pp. 314–320, 2010.

[43] D. Insa, J. Silva, and S. Tamarin, “Using the words/leafs ratio in the DOM tree for content extraction,” *J. Log. Algebr. Program.*, vol. 82, no. 8, pp. 311–325, 2013.

[44] H. Y. Kao, J. M. Ho, and M. S. Chen, “WISDOM: Web Intrapage Informative Structure Mining based on Document Object Model,” *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 5, pp. 614–627, 2005.

[45] Y. Li and J. Yang, “A novel method to extract informative blocks from web pages,” *IJCAI Int. Jt. Conf. Artif. Intell.*, pp. 536–539, 2009.

[46] S.-H. Lin and J.-M. Ho, “Discovering informative content blocks from Web documents,” *Proc. eighth ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD ’02*, p. 588, 2002.

[47] Q. Liu, M. Shao, L. Wu, G. Zhao, G. Fan, and J. Li, “Main content extraction from web pages based on node characteristics,” *J. Comput. Sci. Eng.*, vol. 11, no. 2, pp. 39–48, 2017.

[48] S. López, J. Silva, and D. Insa, “Using the DOM Tree for Content Extraction,” *Electron. Proc.*
Theor. Comput. Sci., vol. 98, pp. 46–59, Oct. 2012.

[49] D. Pan, S. Qiu, and D. Yin, “Web page content extraction method based on link density and statistic,” 2008 Int. Conf. Wirel. Commun. Netw. Mob. Comput. WiCOM 2008, 2008.

[50] P. A. R. Qureshi and N. Memon, “Hybrid model of content extraction,” J. Comput. Syst. Sci., vol. 78, no. 4, pp. 1248–1257, 2012.

[51] P. Rohilla and O. Sharma, “Web Content Mining : An Implementation on Social Websites,” Int. J. Adv. Res. Comput. Commun. Eng., vol. 4, no. 7, pp. 108–111, 2015.

[52] Y. F. Tseng and H. Y. Kao, “The mining and extraction of primary informative blocks and data objects from systematic Web pages,” Proc. - 2006 IEEE/WIC/ACM Int. Conf. Web Intell. (WI 2006 Main Conf. Proceedings), WI’06, pp. 370–373, 2007.

[53] B. Zhang and X. F. Wang, “Content extraction from Chinese web page based on title and content dependency tree,” J. China Univ. Posts Telecommun., vol. 19, no. SUPPL. 2, p. 147–151, 2012.

[54] C. H. Huang, P. Y. Yen, Y. C. Hung, T. R. Chuang, and H. M. Lee, “Enhancing entropy-based informative block identification using block preclustering technology,” Conf. Proc. - IEEE Int. Conf. Syst. Man Cybern., vol. 3, pp. 2640–2645, 2007.

[55] H. Y. Kao, S. H. Lin, J. M. Ho, and M. S. Chen, “Mining Web Informative Structures and Contents Based on Entropy Analysis,” IEEE Trans. Knowl. Data Eng., vol. 16, no. 1, pp. 41–55, 2004.

[56] S. Ajoudanian and M. D. Jazi, “Deep Web Content Mining,” World Acad. Sci. Eng. Technol., vol. 49, no. 1, pp. 501–505, 2009.

[57] X. Huang et al., “Web Content Extraction Using Clustering with Web Structure,” Int. Symp. Neural Networks, vol. 10261, no. ISSN 2017, pp. 95–103, 2017.

[58] M. Kakol, R. Nielek, and A. Wierzbicki, “Understanding and predicting Web content credibility using the Content Credibility Corpus,” Inf. Process. Manag., vol. 53, no. 5, pp. 1043–1061, 2017.

[59] C. Menaka and N. Nagadeepa, “Clustering and Labeling of Images under Web Content Mining,” Int. J. Adv. Res. Comput. Sci., vol. 7, no. 5, pp. 75–79, 2016.

[60] T. Weninger, W. H. Hsu, and J. Han, “CETR: content extraction via tag ratios,” Proc. 19th Int. Conf. World wide web, pp. 971–980, 2010.

[61] J. Smižanský, J. Pokorny, and J. Smizansky, “Page content rank: an approach to the web content mining,” Proceeding IADIS Appl. Comput., pp. 289–296, 2005.

[62] S. Debnath, P. Mitra, N. Pal, and C. L. Giles, “Automatic Identification of Informative Sections of Web-pages,” IEEE Trans., vol. 17, no. i, pp. 1–35, 2005.

[63] E. S. Laber et al., “A fast and simple method for extracting relevant content from news webpages,” Proceeding 18th ACM Conf. Inf. Knowl. Manag. - CIKM ’09, p. 1685, 2009.

[64] D. Q. Nguyen, D. Q. Nguyen, S. B. Pham, and T. D. Bui, “A fast template-based approach to automatically identify primary text content of a web page,” KSE 2009 - 1st Int. Conf. Knowl. Syst. Eng., pp. 232–236, 2009.

[65] Q. Peng, Q. Wang, Y. Li, J. Zhang, and Y. Hao, “Content extraction from chinese web pages based on punctuations distribution,” Proc. - 2012 Int. Conf. Comput. Sci. Serv. Syst. CSSS 2012, pp. 1351–1355, 2012.

[66] S. M. AL-Ghuribi and S. Alshomrani, “Bi-languages Mining Algorithm for Extraction Useful Web Contents (BiLex),” Arab. J. Sci. Eng., vol. 40, no. 2, pp. 501–518, 2015.

[67] C. Wang, J. Lu, and G. Zhang, “Mining key information of web pages: A method and its application,” Expert Syst. Appl., vol. 33, no. 2, pp. 425–433, 2007.

[68] T. Gottron, “Content code blurring: A new approach to Content Extraction,” Proc. - Int. Work. Database Expert Syst. Appl. DEXA, pp. 29–33, 2008.

[69] J. Arias, K. Deschacht, and M. F. Moens, “Language Independent Content Extraction from Web Pages,” DIR’09 Proc. 9th Dutch-Belgian Inf. Retr. Work., pp. 50–55, 2009.

[70] K. T. Sambanthan, and S. S. Dhenakaran, “Optimized Web Content Mining,” Life Sci. J., vol.
14, no. 2, pp. 41–47, 2017.

[71] M. Schedl, G. Widmer, P. Knees, and T. Pohle, “A music information system automatically generated via Web content mining techniques,” Inf. Process. Manag., vol. 47, no. 3, pp. 426–439, 2011.

[72] P. Wang, C. Sanín, and E. Szczербicki, “Introducing the concept of decisional DNA-Based web content mining,” Cybern. Syst., vol. 43, no. 2, pp. 136–142, 2012.

[73] P. Sivakumar, “Effectual Web Content Mining using Noise Removal from Web Pages,” Wirel. Pers. Commun., vol. 84, no. 1, pp. 99–121, 2015.

[74] M. Sethi and G. Gaur, “Web Mining Techniques for Extraction of News,” Int. J. Adv. Eng. Technol. Manag. Appl. Sci., vol. 2, no. 4, pp. 34–47, 2015.

[75] Y. C. Wu, “Language independent web news extraction system based on text detection framework,” Inf. Sci. (Ny.), vol. 342, pp. 132–149, 2016.

[76] T. B. N. Hoang and J. Mothe, “Location extraction from tweets,” Inf. Process. Manag., vol. 54, no. 2, pp. 129–144, 2018.

[77] N. R. Kapadia and K. Patel, “Web Content Mining Techniques – a Comprehensive Survey,” Int. J. Res. Eng. Appl. Sci., vol. 2, no. 2, pp. 1869–1877, 2012.

[78] Omolehin, J. O. and Oyelade, O. J. and Oladipupo, O. O. (2005) Application of Fizzy Logic in Decision Making on Student’s academic performance. Bulletin of Pure and Applied Sciences, 24 (2). pp. 381-387.

[79] Daramola, Olawande and Okuboyejo, Sena and Oladipupo, O. O. and Ayo, C. K. (2015) A Cloud-based Framework for e-Empowerment of Businesses in the Nigerian Informal Sector. In: 2nd Covenant University Conference on African Development Issues (CU-ICADI), 11th - 13th May, 2015, Africa Leadership Development Center, Covenant University, Ota, Nigeria.