University of Sindh Journal of Information and Communication Technology (USJICT)
Volume 9, Issue 2, 2025
ISSN-E: 2523-1235, ISSN-P: 2521-5582
Website: http://sujo.usindh.edu.pk/index.php/USJICT/

Web Content Mining Concepts Techniques and a Comparative Methodology for Modern Web Data Extraction

Ali Hassan Sial *, Zubair Sajid, Muhammad Tahir

Department of Computer Science, Faculty of Engineering, Science and Technology (FEST), Iqra University, Karachi
E-mails of authors: ali.hassan01@iqra.edu.pk[1], zubair.sajid@iqra.edu.pk[2], muhammad.tahir01@iqra.edu.pk[3]

Abstract: Rapid development of the World Wide Web (WWW) has led to vast amount of data having diverse structures, both structured, semi-structured and unstructured formats. The challenge has grown critical in the research field to be able to extract meaning and implications from such data. 1.5 Web Content Mining is an important process of applying data mining techniques, text mining and artificial intelligence techniques to convert the unstructured web documents to structured knowledge. This paper summarizes the concepts, techniques and tools used for web content mining, discusses the most recent achievements (with the aid of machine learning and natural language processing). This study is different from the survey metric-based approaches in that it presents a comparative analysis of classical and AI based mining methods, and evaluates them according to their efficiency and their benefits of being accurate and scalable. To implement the proposed lightweight framework, three web content mining tools are evaluated that are widely used in the Web, namely Web Info Extractor, Mozenda and Screen Scraper. In general, the results have broad implications regarding the capabilities and limitations of the existing tools and the better performance and versatility of the new tools based on AI for extraction accuracy. The study results can be of value to the researchers and practitioners involved in building the effective Web content mining systems for the real-world applications like, e-commerce and Cyber security.

Keywords: Web Content Mining, Web Data Extraction, Text Mining, Natural Language Processing, Machine Learning, Web Mining Tools, Comparative Analysis, Information Extraction

I. Introduction

A. Background and Motivation

The World Wide Web (WWW) is the largest and most publicly available repository of information available, creating large amounts of structured, semi-structured and unstructured data. The information space is constantly growing and extremely diverse, including web pages, online reports, social media posts and dynamically created documents. This expansion fosters knowledge sharing, digital innovation, but also poses major challenges with regard to knowledge localization, extraction and organizing, which must be done efficiently in order to retrieve relevant information. Traditional information retrieval manual approaches are mainly based on keyword matching and pre-designed rules. These techniques are becoming more and more fruitless with the evolution of current Web content into evolving and dynamic structures and the existence of semantic ambiguity. Accordingly, there is a need for automated technique that can take care of huge amount of varied data sources to convert the data that is shared as vague facts into meaningful knowledge [1].

B. Web Content Mining and Its Evolution

Generally, web mining can be divided into three categories: web structure mining, web usage mining and web content mining. Of those, Web Content Mining is the special case concerns the extraction of any useful information from the contents of any web document without considering its links or how it's used. It is essential in applications like search engines, recommendation systems, business intelligence, cybersecurity analysis etc. The initial web content mining systems relied on rules-based and/or statistical methods. For static web pages and web pages with a well-structured layout, techniques like pattern matching and text frequency analysis proved to be helpful. These methods, however, have drawbacks in the case of large-scale, dynamic, noisy data spaces [2]. In recent years machine learning and natural language processing have led to great developments in the adaptability and accurate understanding of a web content mining system's semantic comprehension of web content. The models that can be implemented with this AI can learn features automatically, interpret them in context, and have a higher chance of catching the right features within a wide variety of web formats [3, 4].

C. Limitations of Existing Studies

Although many improvements have been made, there remain some gaps in the literature. Many studies are focused on introducing one or a few tools and algorithms, but do not include systematic performance comparison. There are also those who circulate reviews on survey methods, which are descriptions of techniques without empirical evidence or practical evaluation. Furthermore, comparative analysis of the traditional web mining methods and the more modern ones based on the use of AI are not often pursued or discussed [5]. Another constraint is to evaluate web contents mining tools.

Discuss widely-used tools instead of measurable characteristics, such as extraction accuracy, processing speeds, and scalability. Such a difference makes real-world deployment with solutions a challenge, especially for data-intensive applications like e-commerce or cybersecurity [6].

D. Research Objectives and Scope

In order to fill these gaps, in this paper, a systematic and comparative study on web content mining techniques and tools is introduced. The paper also discusses basic research concepts, as well as a simple research method to compare and assess traditional mining methods and artificial intelligence for mining by analyzing these methods using common performance measures. The method focuses on real-world and repeatable techniques and concepts instead of the theoretical aspects. This work deals with structured, semi-structured and unstructured Web data, and focuses on content extraction techniques and the most popular Web mining tools. The assessment process is to enable comparative assessment with application domains or mining paradigms.

E. Paper Contributions

The following are the key contributions to this paper:

- Offers a reconsideration in web content mining methods emphasizing an evolution from standard techniques to AI based.
- Mentions a comparative research approach to assess web content mining techniques and tools.
- Incorporates metrics of performance such as accuracy, efficiency and scale.
- Compares the work of majority of the commonly used web content mining tools with the proposed methodology.
- Illustrates how web content mining can be applicable in web data extraction systems, including e-commerce and cybersecurity systems.

F. Paper Organization

The discussion presented in the rest of this paper follows this organization:

- In **Section II**, related works are reviewed with focus on the recent studies on Web content mining using Artificial Intelligence technologies.
- The proposed research methodology and extraction process is presented in **Section III**.
- **Section IV** discusses major approaches to web content mining.
- Comparative results and discussion are presented in **Section V**.
- The final **Section VI** brings the paper to a close and points out future research directions.

II. RELATED WORK

A recent development in web content mining has been moving away from rule-based and statistical methods to using intelligent, learning-based methods. Gasparito's driving force for this shift is that web data is now more complex, involving dynamic content, multimedia components, and semantically-charged context. Research articles focused on the utilization of machine learning and natural language processing for enhancing the accuracy and adaptability of the extraction process were highlighted, as maintained by the scholars in their research activities between years 2022 and 2025. Research publications published within the period of 2022-2025 give due emphasis on the increased accuracy and adaptability in extraction using machine learning and natural language processing. In [7] Sharma et al. gave an extensive survey of machine learning and NLP based web content mining techniques. Their research show that supervised and unsupervised learning models achieve better performance than those based on keywords extraction which are mainly used for traditional web content dealing with structured web content. Their study however, was mainly about algorithmic classification, and has not considered the practical performance of Web mining tools in real environment. In the work developed by *Ferrara et al*. [8] they have looked into deep learning methods for large-scale Web data extraction. The research they had conducted showed that neural architectures, particularly transformers, have significantly boosted semantic understanding and noise resistance. The authors' conclusions were impressive on benchmark datasets, but the authors did not make a comparison to other methodologies used to extract information, so it is hard to see the impact the results would have on practitioners looking for deployable information extraction tools. *Zhang and Liu* [9] examined the AI-powered Web Data Mining Systems and discussed the major issues of Web scalability, Data privacy, and interpretability of models. Although the methodologies of AI can be more accurate, higher demand is required for computational power and focus on tuning AI models.

Heavy stuff removal, on the other hand tends to use traditional tools to fit the low overhead. This disclaimer suggests that there is a need for equity frameworks for evaluation that take into account performance and usability. The authors of Verma, *Singh et al.* [10] presented a scalable web content mining framework based on transformer architecture of language models. They made significant advances in contextual extraction and text support for multiple languages. The study could however, be considered as model-centric and could be oriented towards integration with the current web mining tools and pipelines that are used by industry. Recently there has been increased interest in tool-oriented evaluations. *Alshammari and Alenezi* [11] made a comparative study between Web scrapping tools for structured data extraction versus unstructured data extraction. They evaluated the tools based on their usability and their support for data format, but accuracy and scalability were not reported. This restraint emphasizes the importance of using a combined, unified evaluation methodology based not only on functional and performance-based factors. As the application point of view is concerned, *Jin and Lin* [12] showed that web content mining is applicable to cyber security by analyzing the web log. Their

findings clearly showed how effective mining methods are in identifying anomalies and threats for security. Recently, web content mining techniques for e-commerce analytics, sentiment analysis and recommendation systems were discussed adding to the current trend to find more reliable and scalable mechanisms for the extraction of information [13]. As a summary, the existing studies bring useful insights about the algorithms, models and applications of web content mining. Most, however, concentrate on individual elements of the performance of the model or particular features of the tool. Traditional evaluation methods, AI-based methods, and practical tools are still separate from each other and lack comparative evaluations. In this paper, this gap will be filled by introducing a research methodology based on a unified evaluation of web content mining techniques and tools and applying their performance metrics.

III. RESEARCH METHODOLOGY

A. Methodological Design

The methodology used in this research is comparative, and the approach taken are evaluation oriented. The aim is not to showcase a new algorithm, but to provide a framework to compare the what can be done with centuries-old mining techniques to newer ones using AI. The method itself should simulate the real-world case of web mining using various data sources, content structures and real constraints during the process of extracting a web content. Its sessions are geared towards a process that can be repeated, easily understood, and measurable.

B. Methodology Dataflow Overview

The proposed methodology just has to be followed in sequential steps starting with acquiring web data and then moving on to performance-based evaluation and analysis. The different components of the pipeline guarantee that the different web mining methods and tools are consistent and comparable.

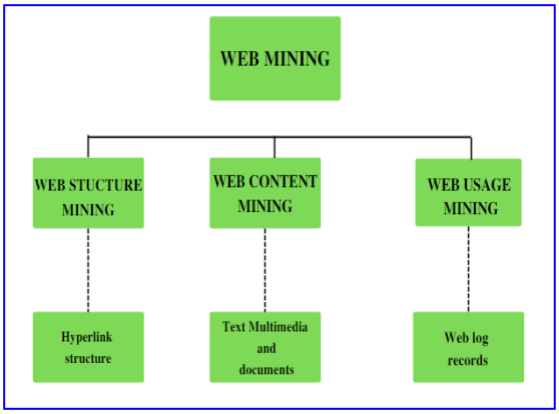


Figure 1. Dataflow of the Proposed Research Methodology.

The information stream of the intended approach to research is depicted in **Figure 1**. It includes the following steps: It begins with heterogeneous web data sources - structured, semi-structured and unstructured data such as HTML pages, tables and textual documents. They can be obtained via web crawler and web crawling technologies with the means of automatically and consistently collecting data from them. All the data gathered is then sent over to the pre-processing step of the pipeline, which eliminates some of the irrelevant components, like ads, scripts and noisy markup. The data is normalized and structured through text parsing to enhance data quality and relevance. The preprocessed data is then segmented into structured, semi-structured and unstructured data types. At this classification step, suitable mining methods for the different types of data can be selected: For structured content, traditional rule-based extraction methods are used, and for the unstructured content, methods like NLP and ML might be used. At the same time, three popular Web Content Mining (WCM) software, namely, Web Info Extractor, Mozenda, and Screen Scraper are used on the same sets of data. At the same time, WCM tools such as Web Info Extractor, Mozenda, and Screen Scraper are used on the same sets of data. This leads to a fair comparison between tools and techniques to be executed in parallel. These extracted outputs are then evaluated on some given set of performance criteria like accuracy, processing time, expandability and acceptability. Finally, the outcome of said results is compared and conclusions are made on the effectiveness of the technique and answers are given to the question(s): does suit for real-life applications.

C. Data Sources and Collection

Ethical compliance and the reproducibility of experiments is performed with data sources that are publicly accessible web pages. The data consists of various formats such as structured tables, semi-structured HTML data and unstructured textual data. Content is retrieved automatically by the crawling and the integrity of any document format is kept by automated methods [14].

D. Data Preprocessing

The preprocessing is carried out to improve the accuracy of the extraction. This involves eliminating unnecessarily long elements, normalizing textual data and parsing document structures. In the case of semi-structured data, hierarchical relationships between the tags are preserved to maintain contextual relationships. This step is for reducing noise and to prepare the data for effective mining [15].

E. Mining Techniques and Tools

The methodology covers both classical methods of extraction and a developed one using artificial intelligence. There are two approaches for structured information: rule-based and wrapper-based methods, and machine learning and NLP-based methods for unstructured information. In order to obtain an unbiased evaluation, the selected web content mining tools are evaluated under similar conditions [16, 17].

*F. **Evaluation Metrics***

The following measure are used to obtain the objective assessment:

- **Extraction Accuracy**: the accuracy of the information retrieved.
- **Processing Efficiency**: amount of time it takes to extract from the product.
- **Scalability and Usability**: increase in the amount of data handled. In many cases, the ability to configure and operate easily determines the usability.

These measures provide an unbiased assessment of technical capabilities and application [18].

G. Methodological Significance

The proposed methodology, method of bridging the gap between theoretical studies and deployment in practice. Comparative evaluation, combined with real-world tools, provides a structured evaluation method for web content mining solutions. To enable researchers and practitioners in the fields of big data on the web to understand and act on it.

IV. APPROACHES OF WEB CONTENT MINING

Web content mining is aimed at extracting meaningful information from web documents based on the actual content of them. Three types of content mining can be determined by the structure of web data: unstructured, structured and semi-structured mining. The classification is suitable for the research methodology that is presented and proposed which depends on the format and complexity of the data that will be leveraged and extracted through different methods [34 - 40].

A. Unstructured Web Content Mining

The majority of the information in an unstructured web is free-text content like articles, blogs, reviews, and social media posts. The data doesn't have a standardized schema, so meaningful information cannot be easily obtained without using sophisticated text processing and semantic analysis techniques. In recent approaches, there are more investigations focused on Utilization of Natural Language Processing and Machine Learning Models for improving extraction accuracy and contextual understanding [19].

Information Extraction: Information is concerned with extracting the relevant entities, relationships and attributes from unstructured textual content. The traditional approaches that were used were pattern matching and rule-based systems, but they suffered greatly when it came to generalizing in order to be able to handle less commonly used languages and ambiguities were even harder to deal with. Recent developments in Named Entity Recognition and semantic annotation have been used by AI techniques to make the extraction process more accurate and scalable [20].

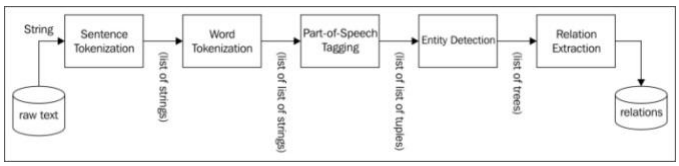
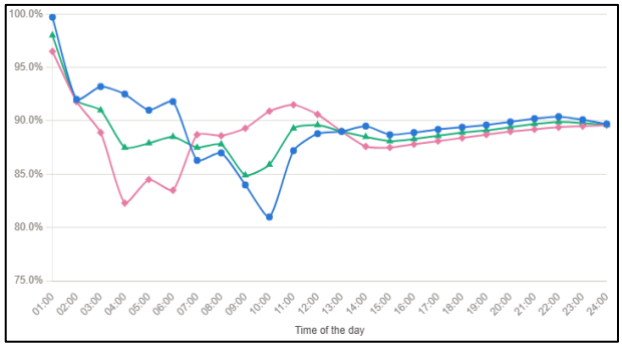


Figure 2. Illustrates the Block Diagram of the Information Extraction Process.

Topic Tracking: which involve the study of the document stream to identify the content, relevant to the existing and/or programmer-defined topics. It's common in news aggregators and news recommenders. Previous approaches use keyword frequency analysis [17] while current approaches use probabilistic models and embeddings based on contextual information to enhance the faithfulness and precision of topics [21].

Figure 3. Presents the Graphical Representation of Topic Tracking.

Text Summarization: When two large documents are



reduced into concise one, without losing any key information, it is summarized via Text summarization. Extractive summarization extracts key sentences from the original text while abstractive summarization creates a model-based semantically valuable summary with NLP models. The quality of abstractive summarization of long web documents has been substantially improved in recent years by engineering improvements to transformer-based systems, especially for web documents [22].

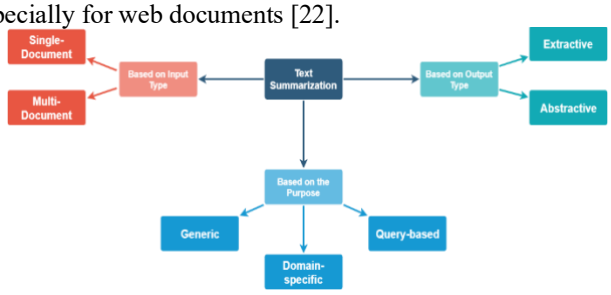


Figure 4. Shows Text Summarization and Classification.

Categorization and Clustering: Categorization is the technique of assigning documents to classes, where classes are used beforehand for document classification, Clustering is the technique that groups documents in clusters without using any predefined classes for document classification.

The techniques can be used to organize and retrieve web content efficiently. A large-scale web dataset is often tackled

with the help of the machine learning-made classifiers and unsupervised clustering algorithms [23].

Information Visualization: Information visualization helps in exploring big collections of documents via visual forms like maps, hierarchies and networks. Visualization techniques have been found to support user interaction in the discovery of user patterns and relationships within collections of web contents [24].

B. Structured Web Content Mining

Structured Web Content Mining refers to structured data, typically in the form of a table, list or database, which is found on websites. Since it is a structured document, extraction of structured data is relatively easy and often can be done by applying rule-based parsing methods.

Web Crawlers: These are the programs that systematically trawl web pages to gather and catalog the contents. Crawlers are important for the acquisition of structured data at scale, and have become the cornerstone of search engines and mining applications [25].

Web Page Content Mining: This is an approach that looks at the information that needs to be extracted from a web page that has a structured form and is ranked according to a web page ranking system. Page ranking algorithms help in better identification of authoritative and relevant pages, which help in boosting the efficiencies in their extraction [26].

Wrapper Generation: Wrapper Generation automatically extracts structured information using structured patterns found in web pages. Wrappers are useful for static layouts, but are not scalable as content changes rapidly over time on the pages [27].

C. Semi-Structured Web Content Mining

Semi-structured data is not structured or unstructured, but have a format that includes HTML docs, XML docs and JSON docs. Such data may not have a rigid schema, but it does have implicit structural tags which makes it easy to extract [33].

The Object Exchange Model: The Object Exchange Model is used for semi-structured information to represent the data as objects and to allow the efficient traversal and retrieval of hierarchical relationships. OEM is a framework to support flexible queries and structural interpretation of web data [28].

Web Data Extraction and Top-Down Approaches: These methods are techniques that utilize the web to extract and convert its semi-structured data into a structured format for analyzing the data. Top-down extraction methods iteratively break a complex object into parts that are simpler at the end for the required info to be extracted. These methods are aligned with proposed methodology since they allow to extract in a scalable and format-aware manner [29].

D. Alignment with the Proposed Methodology

The various approaches in web content mining directly contribute to the proposed research methodology.

Data is organized, semi-organized and un-organized, all fair and consistent and mapped to other techniques/tools. This alignment can help measure the effectiveness of traditional and AI-based techniques and evaluate some of the essential aspects of the techniques, such as accuracy, efficiency, and scalability.

V. RESULTS AND DISCUSSION

In this section, the comparative study of selected web content mining tools is shown based on the proposed methodology for web content mining research. An aim of this evaluation is not just listing the features, but to analyses the performance of the tools in terms of extraction capability, efficiency, scalability and usability. A set of identical data sets were used for all tools, to ensure fairness and consistency.

A. Experimental Setup

The items were evaluated on public web pages which had structured tables, semi-structured HTML and textual data in an unstructured format. Selected tools were set to its default extraction tools to match the common practice that occurs in the real world, Web Info Extractor, Mozenda and Screen Scraper. Observations were conducted of the performance with a number of extraction tasks and qualitative observations were conducted for the accuracy, processing behavior and ease of use. The test criteria adopted are those specified in (Section III).

B. Comparative Results of Web Content Mining Tools

To compare and analyze these tools and check the performance of all the tools. The evaluation was based on whether each of these tools could manage structured and unstructured data; processing efficiency, scalability, and usability. These criteria had been selected for their ability to represent deployments, which would be realistic, and not merely capabilities. These findings reveal significant variations in tool performance when compared with the others that were tested. Cloud-based tools with automatic support and configuration are more compatible for dealing with a variety of large web data. Traditional extraction tools, in contrast, are good for extracting structured data sources, and do less well for unstructured data sources. There is also variance in methods for usability: graphical interfaces tend to be extremely cumbersome and user-friendly, while a script-based interfaces would have more manual labor and technical expertise for users to enter. The result that can be compared suggests indeed a trade-off in terms of performance – usability problems. The advanced tools will be more accurate in their extraction and more scalable but the simple will be helpful in controlled environments where they'll be easier to set up and less computing power. The summary of the observation when compared to the criterion of assessment is presented in **Table 1** below as follows:

C. Discussion of Results

The outcome shows that Mozenda is performing well for both structured and unstructured data extraction. The cloud-based architecture ensures higher scalability and quicker processing, thus making it suitable for large-scale applications. However, the process to store it is moderate technology. Web Info Extractor can extract very well for structured content, and has high usability through its graphical configuration interface. Of course, it worked well for smaller

to mid-sized jobs, but it isn't a strong option for unstructured information compared to the work of artificial intelligence. Screen Scraper is moderately effective, albeit with some flexibility problems and problems with unstructured data and scalability. They are very cumbersome to configure and script, therefore decreasing ease of use and usability in dynamically changing web environments. Overall, the findings indicate that AI-powered and cloud-based capabilities outperform traditional extraction systems in terms of flexibility and such scalability. But for simple and light extraction, there are simpler tools which are still relevant.

D. Alignment with Proposed Methodology

Evaluation based on comparative analysis shows the effectiveness of the proposed research methodology. The methodology is used using a consistent set of datasets and a consistent set of assessment criteria so that an objective comparison of tools and extraction approaches can be made. This within the environment of a structured evaluation reflects both advantages and drawbacks of performance, scalability and usability to help inform choices about the tools to use for a particular application.

E. Practical Implications

The results highlight that there is no best practice tool for every situation. AI tools and scalable tools work better for large-scale and dynamic environments, like e-commerce sites or for cybersecurity monitoring. However, traditional tools are still applicable for smaller data sets where simplicity and a low overhead is desirable. However, in smaller data sets where simplicity and minimizing overhead is paramount, traditional tools are still viable options.

VI. CONCLUSION AND FUTURE WORK

As there is a lot of web content along with many challenges in extracting meaningful information from a heterogeneous data that are approaching the extraction, this paper, presents a structured and comparative study of web content mining concepts, techniques and tools. The study discussed the history of web content mining and covered the structured, semi-structured, and unstructured data in the website. It showed how the web content mining is evolving from old rule-based approaches to recent AI-based methods with machine learning and natural language processing. Compared to the purely descriptive surveys, it was not only a systematic work based on performance metrics that can be consistently used for evaluating web content mining techniques and tools, but also it was an introduction of the light-weight research methodology. For more complex and dynamic web contents, AI assisted tools and AI powered cloud services tended to be more flexible, more accurate in extracting content and easier to scale up. However, traditional approaches continue to be important in certain controlled settings in which convenience and a lack of demand for computation are essential. These results show the value of the methodology used in practice and indicate the objectives mentioned in the abstract study are valid. The results also confirm the importance of balancing the different aspects of performance, usability and scalability in order to build an effective content mining application for the Web, instead of using only one of the extraction strategies. Based on the data types and the application needs, the proposed architecture is specifically designed, which will benefit the decision-making process of the researchers and practitioners in various domains such as ecommerce analytics, cybersecurity monitoring, etc.

Further work: The work can be expanded for future scenarios by applying quantitative benchmarking on larger datasets and across diverse scenarios and using advanced deep learning and transformer-based models in their proposed methodology. Further research investigating the integration of techniques and ideas for privacy aware mining and automated adaptation in dynamic web environments is also promising.

Acknowledgment: The authors are grateful to Iqra University, Karachi for its research-friendly environment. The authors would also like to thank anonymous reviewers for their valuable comments and sound suggestions to enhance the quality of this paper.

Conflict of Interest: The authors state that they have no conflict of interest for the publication of this research paper.

REFERENCES

[1] A. Sharma, R. Kumar, and S. Bansal, "A survey on web content mining techniques using machine learning and natural language processing," IEEE Access, vol. 10, pp. 114233–114250, 2022.

[2] M. J. H. Mughal, "Data mining: Web data mining techniques, tools and algorithms: An overview," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 6, pp. 1–10, 2018.

[3] M. A. Ferrara, G. D. Mauro, and S. Greco, "Deep learning approaches for large-scale web data extraction," ACM Computing Surveys, vol. 55, no. 6, pp. 1–36, 2023.

[4] S. Zhang and H. Liu, "AI-driven web data mining: Challenges, techniques, and future directions," Future Generation Computer Systems, vol. 145, pp. 1–14, 2024.

[5] K. Verma and P. Singh, "Scalable web content mining using transformer-based language models," Expert Systems with Applications, vol. 235, 2025.

[6] T. Alshammari and M. Alenezi, "Comparative analysis of web scraping tools for structured and unstructured data extraction," Journal of Big Data, vol. 11, no. 1, 2024.

[7] A. Sharma, R. Kumar, and S. Bansal, "A survey on web content mining techniques using machine learning and natural language processing," IEEE Access, vol. 10, pp. 114233–114250, 2022.

[8] M. A. Ferrara, G. D. Mauro, and S. Greco, "Deep learning approaches for large-scale web data extraction," ACM Computing Surveys, vol. 55, no. 6, pp. 1–36, 2023.

[9] S. Zhang and H. Liu, "AI-driven web data mining: Challenges, techniques, and future directions," Future Generation Computer Systems, vol. 145, pp. 1–14, 2024.

[10] K. Verma and P. Singh, "Scalable web content mining using transformer-based language models," Expert Systems with Applications, vol. 235, 2025.

[11] T. Alshammari and M. Alenezi, "Comparative analysis of web scraping tools for structured and unstructured data extraction," Journal of Big Data, vol. 11, no. 1, 2024.

[12] J. Jin and X. Lin, "Web log analysis and security assessment method based on data mining," Comput. Intell. Neurosci., vol. 2022, Article ID 8485014, 2022.

[13] R. Gupta and P. Malhotra, "Web content mining for e-commerce analytics and recommendation systems," IEEE Transactions on Computational Social Systems, vol. 10, no. 2, pp. 312–324, 2023.

[14] H. Liu and B. Lang, "Web crawling and scraping techniques for large-scale data acquisition," IEEE Internet Computing, vol. 26, no. 4, pp. 45–53, 2022.

[15] Y. Kim and J. Lee, "Preprocessing techniques for unstructured web text mining," Knowledge-Based Systems, vol. 248, 2022.

[16] P. Singh and A. Kaur, "Classification of web data using hybrid mining techniques," Expert Systems with Applications, vol. 210, 2023.

[17] M. Oliveira and R. Torres, "Evaluating web content mining tools in heterogeneous data environments," Information Processing & Management, vol. 60, no. 1, 2023.

[18] L. Chen and W. Zhao, "Performance metrics for scalable web data extraction systems," Journal of Big Data Analytics, vol. 6, no. 2, 2024.

[19] S. Zhang and H. Liu, "AI-driven web data mining: Challenges, techniques, and future directions," Future Generation Computer Systems, vol. 145, pp. 1–14, 2024.

[20] M. A. Ferrara, G. D. Mauro, and S. Greco, "Deep learning approaches for large-scale web data extraction," ACM Computing Surveys, vol. 55, no. 6, pp. 1–36, 2023.

[21] A. Sharma, R. Kumar, and S. Bansal, "A survey on web content mining techniques using machine learning and NLP," IEEE Access, vol. 10, pp. 114233–114250, 2022.

[22] K. Verma and P. Singh, "Scalable web content mining using transformer-based language models," Expert Systems with Applications, vol. 235, 2025.

[23] P. Singh and A. Kaur, "Classification of web data using hybrid mining techniques," Expert Systems with Applications, vol. 210, 2023.

[24] R. Gupta and P. Malhotra, "Visualization techniques for large-scale web content analysis," IEEE Computer Graphics and Applications, vol. 43, no. 2, pp. 88–97, 2023.

[25] H. Liu and B. Lang, "Web crawling and scraping techniques for large-scale data acquisition," IEEE Internet Computing, vol. 26, no. 4, pp. 45–53, 2022.

[26] T. Alshammari and M. Alenezi, "Comparative analysis of web scraping tools for structured and unstructured data extraction," Journal of Big Data, vol. 11, no. 1, 2024.

[27] M. Oliveira and R. Torres, "Evaluating wrapper-based extraction systems in dynamic web environments," Information Processing & Management, vol. 60, no. 1, 2023.

[28] Y. Kim and J. Lee, "Semi-structured data modeling and extraction using OEM," Knowledge-Based Systems, vol. 252, 2022.

[29] L. Chen and W. Zhao, "Top-down extraction strategies for scalable web data mining," Journal of Big Data Analytics, vol. 6, no. 2, 2024.

[30] T. Alshammari and M. Alenezi, "Comparative analysis of web scraping tools for structured and unstructured data extraction," Journal of Big Data, vol. 11, no. 1, 2024.

[31] M. Oliveira and R. Torres, "Evaluating web content mining tools in heterogeneous data environments," Information Processing & Management, vol. 60, no. 1, 2023.

[32] L. Chen and W. Zhao, "Performance metrics for scalable web data extraction systems," Journal of Big Data Analytics, vol. 6, no. 2,2024.

[33] H. Gul, S. Jan, I. A. Shah and S. U. Rehman, "A Taxonomy of Text Mining", USJICT, vol. 6, no. 2, 2022.

[34] J. H. Awan, "Exploring Inter-connectivity Link Prediction: An Insight from Social Network Science", USJICT, vol. 8, no. 1,2024.

[35] T. H. Mirjat et al., "Automated Assessment and Learning Framework for Competency-Based Training in TEVT Institutions of Sindh, Pakistan," Spectrum of Engineering Sciences, vol. 3, 2025, pp. 1595–1616.

[36] N. Jawaid, A. Warsi, A. Salam, H. Yaseen, Z. Sajid, E. Ahmed, et al., "Dimensions of Knowledge Graph Reasoning," Spectrum of Engineering Sciences, vol. 3, no. 9, 2025, pp. 1404–1432. [19] A. Wahab, "AI and Machine Learning-Driven Framework for Early Detection and Prevention of Ransomware Attacks in Banking Systems," Policy Research Journal (PRJ), vol. 3, no. 10, Oct. 2025, pp. 751–764.

[37] H. Bux, K. T. Pathan, M. Tahir, Z. Sajid, H. Yaseen, M. Yousuf, et al., "A Context-Aware Learning Framework to Enhance Accessibility for Visually Impaired Students in Higher Education," Spectrum of Engineering Sciences, 2025 (in press / no issue information provided).

[38] Z. Sajid, M. Tahir, H. Bux, A. Salam, C. D'Silva, and I. Hussain, "Robust Real-Time 2D Object Detection Using YOLOv5: Architecture, Training Optimization, and Comparative Evaluation," Spectrum of Engineering Sciences, 2025 (in press).

[39] J. Yang, B. Zhou, M. Zhang, X. Zheng, and X. Liu, "Multi-Source Consistency Deep Learning for Semi-Supervised Operating Condition Recognition in Sucker-Rod Pumping Wells," International Journal of Advanced Computer Science and Applications, vol. 15, no. 12, 2024.

[40] "Behavioral Drivers Influencing Cloud Computing Adoption in Pakistan's Financial Sector: A TPB-Based Empirical Study," Center for Management Science Research Journal, vol. 3, no. 6, 2025, pp. 379–395. [Online]. Available: https://cmsrjournal.com/index.php/Journal/ article/view/495