Exploring Web Mining for Tech Awareness: Analyzing Heterogeneous Online Sources

Shah Hussain Bangash[*], Irfanullah Khan[†], Waqas Ahmad[‡], Asfandyar Ahmed[§], Zabih Ullah Khan[**]

Abstract

This proposed study focuses on utilizing web mining techniques to enhance technology awareness. It delves into diverse online sources, including HTML documents, images, and multimedia, to gain insights into technology updates and research issues. The methodology begins with an assessment of the number of firms with websites within the total population. Surveys are conducted to scrutinize web content, specifically examining firm characteristics and Uniform Resource Locators (URL) coverage. The study emphasizes identifying patterns in URLs to craft an effective search strategy. It also addresses the challenges encountered in web mining, particularly in the context of small firms. It investigates how web mining can drive innovation within firms and improve data analysis processes. Redirecting URLs in large-scale studies is given priority to ensure precise data access and mitigate errors. Furthermore, the study underscores the importance of web mining in managing web page content across firms of various sizes. This research study significantly contributes to augmenting technology awareness by conducting a comprehensive analysis of web mining techniques. It also addresses different technology updates and research issues encountered in various online sources, thereby advancing our understanding of the field and its implications for technological advancements.

Keywords: Web Scraping; Web Mining Techniques; Web Content Mining; Web Usage Mining; Search Strategy; Data Analysis Processes.

Introduction

Web mining is an important technique for automatically discovering and triggering useful information to get the optimal knowledge from the Internet, page contents, and hyperlink structure. Web Mining Corporation to improve the capability of a web search engine by analyzing the web documents and revealing the web pages. The information in terms of web content and web pages includes audios, videos, lists, images,

[*]Department of Computer Science, Iqra National University, Peshawar 25124, Pakistan, onlinesofttech@gmail.com

[††]Department of Computer Science, Iqra National University, Peshawar 25124, Pakistan, irfi0092@gmail.com

[‡]Corresponding Author: Department of Computer Science, Iqra National University, Peshawar 25124, Pakistan, waqas.ahmad@inu.edu.pk

[§]Department of Computer Science, Iqra National University, Peshawar 25124, Pakistan, asfandyarahmed@inu.edu.pk

[**]Department of Computer Science, CECOS University of IT & Emerging Sciences, Peshawar 25124, Pakistan, zabihsoft786@gmail.com

hyperlinks, charts, and tables, which are stored in the database and accessed through the server logs. The web regularly expands as the new data is added to it daily. As a result, there is an increasing demand for sophisticated tools to analyze and extract relevant data from this material. Data mining, artificial intelligence, statistics, informatics, and computational linguistics are just a few technologies that are integrated into web mining research. The study of hidden patterns, trends, and information inside online data that conventional data mining approaches can miss is made possible by this multidisciplinary approach. Web Usage Mining (WUM), Web Content Mining (WCM), and Web Structure Mining (WSM) are the three primary classifications into which web mining techniques fall (Chen et al., 2024). Each lesson focuses on a different area of web data, such as extracting information from web page content, analyzing web architecture, and comprehending user behavior patterns on the web. Web mining has numerous and significant uses in today's digital world.

Web mining techniques provide useful insights for organizations, researchers, and decision-makers by recognizing web items with specified qualities, grouping web pages based on comparable subjects, and uncovering trends in user behavior. Researchers can find new insights, boost the effectiveness of online information retrieval, and improve decision-making processes by utilizing modern data mining techniques on web data (Du et al., 2023).

Web mining involves a few critical phases, including information selection/pre-processing, generalization, and analysis/validation. These procedures are designed to extract significant insights from online data by translating it into a standard format, recognizing user access patterns, and evaluating them against known patterns. This organized technique enables researchers to glean hidden insights from a massive pool of online material, improving their understanding of user behavior and preferences (Das et al., 2023). The difficulties of information overload and relevancy continue to exist in the ever-expanding World Wide Web ecosystem. Search engines play an important role in aiding users in navigating this immense sea of information; nevertheless, the sheer volume of data frequently results in result sets that are cluttered with useless items. Web mining approaches, such as web use mining and content mining, address this issue by offering personalized suggestions and boosting search performance (Du et al., 2023; Govers et al., 2023; Gheisari et al., 2023).

Web mining researchers can enhance user experience and simplify information retrieval by utilizing vast amounts of information on the web, including structured and unstructured data, for predictive analytics, recommendations, and service optimization (Kumar & Kumar, 2021; Kayser et al., 2020).

Literature Review

In the early stage of the Internet, various approaches and techniques can be used while gradually the proposed methods researchers collaborate to improve the optimal methods. Web mining as a survey approach seems to determine the highly hyper-connected deviation in websites, and the fact of low broadband prevents some firms from excluding web mining analysis. The transfer firm website data engaged into significant indicators of innovation proposes the framework innovation ecosystem mentions some approaches. On the other hand, most of the innovation ecosystems are at small-scale levels restricted to the timeliness questionnaire survey of granularity (Lee et al., 2022). Other authors have suggested valuable systems and have wanted to spread awareness of the latest technological improvement in web mining from online sources to access information from heterogeneous complex websites (Mohammadi et al., 2022) have propounded web mining methods for gaining valuable information from the web.

The two important approaches for web mining include accessible agent-based and database approaches. The agent-based approach provides an organized, filtered, and categorized mechanism to user information according to a particular query. It also helps apply the recycling of semi-structured data from web content to the database approach (Mele et al., 2019) have studied issues that arise during the research study in web mining. Sometimes, users' searching becomes challenging in that it does not return the information with the relevant type of query. The high- and low-level details from many search engines struggle to develop a quick response to the user's query and provide them with accurate results.

The e-commerce sites are struggling to attract their clients with new ideas in line with the recommendation of the customer facilities components (Rhayem et al., 2020). The study proposed to survey to see the semantic web impact for an improved variety of techniques-based web mining key information about the current state of WWW. Further, the amount of data processing corporative information decreases the issues of overburdened information and suggests a better way of optimal searching classification and investigation (Schedlbauer et al., 2021) to help a review of earlier studies and what researchers and scientists in web prospecting dealt with.

Web mining discovers the knowledge and patterns for the classification of better techniques in searching quickly (Shi et al., 2017) works on the development of the framework of blockchain technology based web mining, which is systematically reduced and organized for Government of India applications, appearing as hidden pattern problems in the organization of new health techniques and tools. Sometimes, however,

the right government data may be hidden from publication, as data mining is a technique that extracts the information. Blockchain is a new technology that is used to protect from cyber-attacks and access secret information without being permitted to do so (Wu et al., 2021).

It is worth noting that many other researchers who emphasize the application of deep learning techniques to predict the outcomes and recognize the model's efficiency seek to improve web search systems. With the help of a popular system, it is possible to find out the economic status and collect the tweet's covid-19 dataset. The applied semantic relationship framework of web mining and the deep neural network approach for a recommended reputation system is furnished with the "Data Mining Techniques Applied for Automatically Extracting the Optimal Data from Web Documents and Services in the Proposed System". This thus documents how the dynamic structure of the World Wide Web has numerous complexities in web documents.

Challenges in Data Mining

Large Data: The volume of data speedily increasing throughout the world. It is a challenging task to extract meaning from the huge dataset.

Identify the Quality of Data: In the field of data mining, another optimal factor is to find out the quality of data. Most of the researchers struggle and emphasize data because data is the first step to analyzing and training the model.

Data Security and Privacy: It is another issue to protect sensitive information through privacy and security. Security breaches a significant role in each organization to protect the outcomes.

Scalability: Today the world become procurement and advanced speedily. The size of the dataset gradually becomes the highest and most complex to deal with. It is important to manage the data properly and become efficient. According to data mining experts, researchers mentioned most algorithms may not be able to deal with huge amounts of data.

Complex and Unstructured Data: If you study data mining techniques and concepts designed for structured data most researchers use them for unstructured data. It will create issues during the process of unstructured data.

Feature Engineering: Feature engineering techniques help us to determine the relevant features or variables in data. It is an important phase to avoid unnecessary, irrelevant, noisy features in the dataset.

Algorithm Selection: The accurate algorithm selection plays the main role in specifying the problems and finding out the optimal solutions. It is a complex issue to search for the right algorithms and make the combination of classifiers with each other.

Research Methodology

Analyzing and investigating how many firms' websites are running from the total population. The researchers hold an allowance for permission to do a survey study of the web-based content. We consider firm characteristics and coverage Uniform resource locators (URLs) through statistical relationships in the overall population. Sometimes the firm dataset is missing URLs between websites' true and false missing values data provider. Our study outcome, regularities in URL controlling the search strategy. The research is how the researchers and scientists manage the web mining ecosystem innovations difficulties with the observing of the tiny firms.

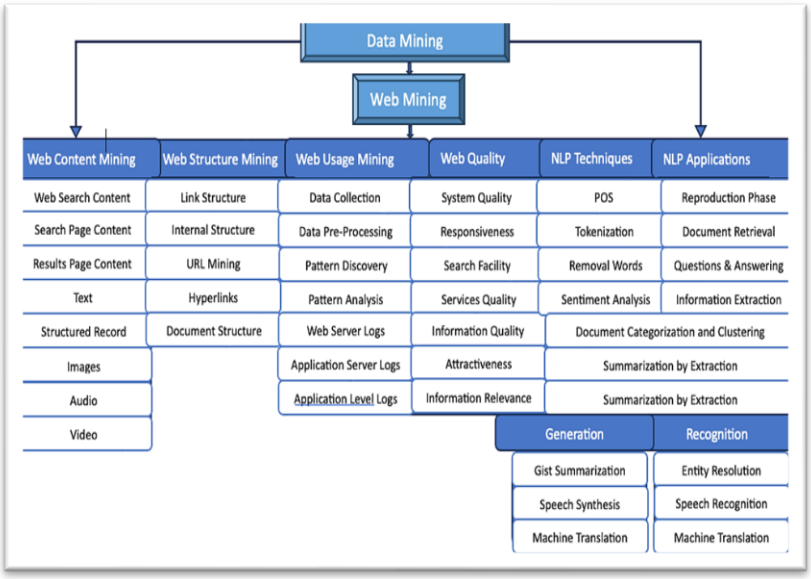

Figure 1: Block Diagram of Data Mining Process data in various stages.

Web mining framework is the intensive firm innovation system to establish a suitable data mining assumption. The conducting redirects URL become harmful determination large-scale web studies in huge datasets. This, therefore, emphasizes our need to reduce the error on the firm's websites, so that we check on the redirects of the URL and assess the data. Mostly, the URL requests are successful for the systematic firm to be redirected. The size of the firms highly recommended web mining to control and manage the webpages of websites. Larger firms have more pages and, consequently, more text available on web pages. Problems in web mining approaches often happen in outliers, while errors in websites.

Data Mining

Data mining is mainly divided into three useful stages, including WCM, WSM, and WUM. Further, these phases are divided into stages, as already mentioned, such that each one is used for a particular purpose. WCM is further divided into three categories including Web search content, Search page Content, and Results page content. The earlier three mining phases are applied or used independently, and other jobs are together, perhaps they include web document links. Web mining is the technology of data mining that aims at retrieving and triggering useful information from web services.

Figure 2: Data Mining Processing Data Various Techniques

As Figure 2 shows, the business process of transforming unprocessed data into actionable and valuable information is known as data mining. Gathering and compiling information from common repositories, including relational, transactional, multimedia, geographic, and data warehouse databases, is known as data mining.

Web Structure Mining: WSM is categorized into three major parts: link structure, internal structure, and URL mining. This type of mining, on the web structure server, accesses a web file which gets attached to the network diagram, and investigates further nodes in it. The link structure connects with other pages of websites internally through the architecture of web mining techniques. URL mining calls for the internal pages of a website and accesses the information.

Web Usage Mining: WUM is categorized into four essential phases: data collection, pre-processing, pattern discovery, and pattern analysis. These methods play significant roles in the techniques of web mining.

Data Collection: The first phase of the research mechanism set for the collection and measurement of important data of interest well-established systematic technique variables. Data collection consists of main parts that answer research questions, evaluate outcomes, and hypotheses tests as well as common all fields of study. It is very significant in collecting appropriate and accurate data (qualitative and quantitative) and integrity of research.

Data Pre-Processing: This is a very important phase to clean the data and reduce the complexities of time as well as improve the accuracy. Before the data preprocessing data will be noisy and redundant, incomplete, inconsistent, missing values, and remove nonvalues. Data preprocessing has some significant features to remove unnecessary features operations like data cleaning, data reduction, data integrity, and data transformation.

Data Cleaning: The process of removing any unwanted data from the dataset and analyzing raw data. After downloading the data set from Kaggle or the UCI repository, it displayed improper results while also creating ambiguity from the given dataset. In case the dataset has noisy and redundant feature values, then it would result in poor accuracies.

Data Reduction: The way of reducing the volume of ambiguity in the dataset. Sometimes data is duplicated and repeat the rows and columns in a dataset. On the other hand, data reduction is a mechanism to remove duplication of features and increase accuracy from the original data.

Data Integrity: These refer to the action of accuracy and security in the validity of the feature dataset to provide consistency. The integrity of solutions provides many ways of resolving and compromising a replicated unaltered data update. It has the responsibility of securing data, reliability, and making sure that the results given are complete and accurate.

Data Transformation: It also includes a process where the form of the feature data structure is evaluated. The stages in the data transformation process encompass data warehousing, data wrangling, data integration, and data migration related to the undergoing data transformation.

Pattern Discovery: The task is to develop and discover the optimal pattern containing the abnormal and periodic pattern for searching engine information from the users' selected query and accessed information websites. Data mining is one of the patterns that are friendly, best, and attractive to discover.

Pattern Analysis: The specific rule of determination pattern regulation market-based analysis through data mining techniques. Mostly, users have difficulty finding the best pattern that is more secure and safe. The rule represents the selection of an appropriate pattern of validation and interesting patterns.

Results

The results appear to be a four-plot representation of the findings of a statistical analysis. The study uses a four-plot arrangement to evaluate model assumptions and data distribution. The Normal Probability Plot compares data points to a straight line, the Histogram shows frequency distribution, and a residuals plot shows residuals from a fitted model. Figure 3 shows the normal probability in the data mining process. Measurement

evaluates the data the normal (Gaussian) distribution model. Data miners estimate the probability that values or events will occur inside a dataset by computing probabilities within the normal distribution curve. Anomaly detection, clustering, and classification are just a few of the data mining tasks that this measurement helps with.

Figure 4 indicates the residuals are shown by the x-axis in the graphic, while the fitted values are represented by the y-axis. The discrepancy between a variable's observed values and its model-predicted values is known as its residual. The values that a model predicts are called fitted values. A scatter plot's point arrangement might show how the variables relate to one another.

The normal probability plot in Figure 5 compares data's cumulative probability with theoretical quantiles of a normal distribution. The straight line representing expected outcomes and blue spots representing actual data. The data exhibits nonlinearity in the middle, resulting in an "S" shape, suggesting skewness or kurtosis, and significant tail deviation, indicating potential outliers or increased variability. The non-normality of the response variable “Loss Severity” may exist due to the loss distribution's skewness, potentially necessitating transformation or alternative statistical methods for modeling.

This kind of plot displays a variable's distribution. The variable's various values are shown by the x-axis, and the frequency or count of data points that fall into each interval is represented by the y-axis. The image's histogram demonstrates how the variable loss severity is skewed to the right, with a greater number of data points falling in the lower value range. Plots that display the relationship between a variable and its order in the data set are known as versus-order plots.

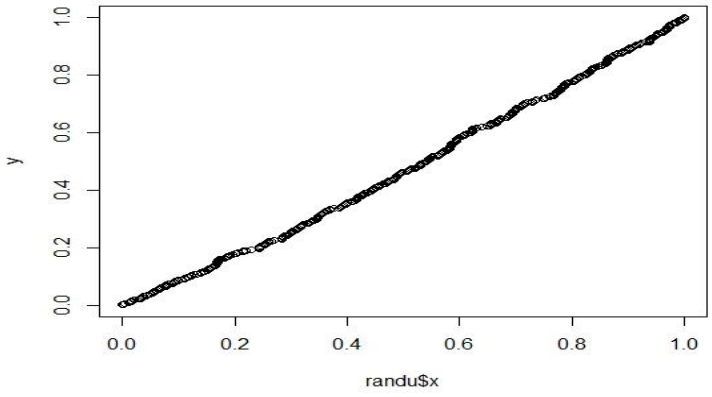

Figure 3: Normal Probability Measurement.

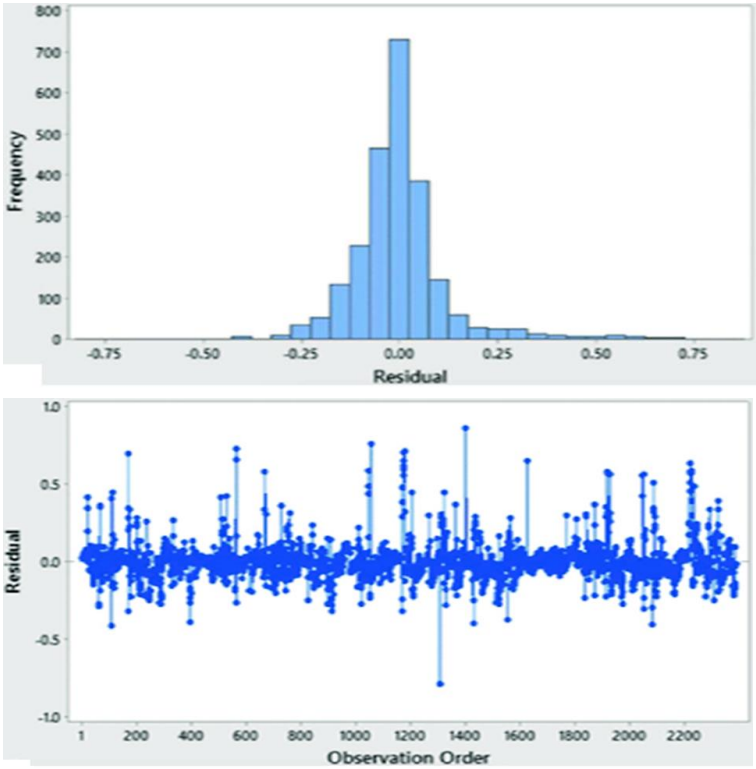

Figure 4: Residuals and observation order regression obtained from the data mining probability.

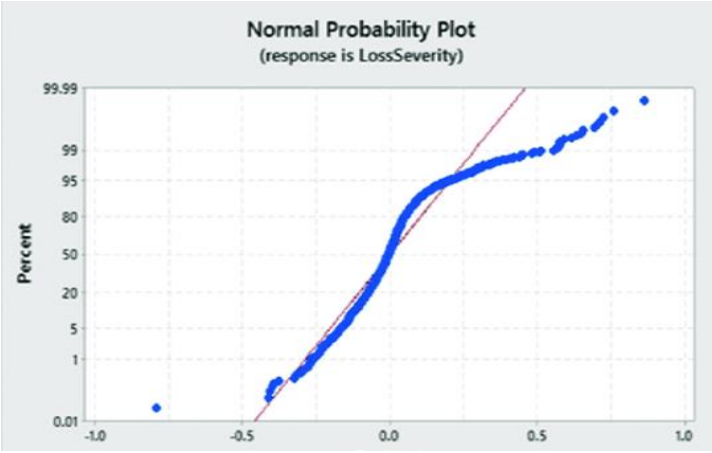

Figure 5: Result of normal probability plot.

Overall, the images imply that the residuals and fitted values have no discernible relationship, and the variable loss severity is not normally distributed. There are no discernible patterns across time, and the variable is skewed to the right. Without more information regarding the data and the analysis that is done, it is challenging to make any firm judgments about the image.

Figure 6 shows the clustering of assembling comparable data points to identify underlying structures and patterns in the dataset. The first step in clustering efficiently is to choose the right algorithm, such as DBSCAN, K-means, or hierarchical clustering, depending on the needs and properties of the data. To maintain uniformity, preprocess the data by addressing missing values, scaling features, or normalizing them. Utilizing the selected algorithm, divide the data into clusters considering cluster centroids and distance metrics. Utilize either internal or external validation measures to assess the clustering results.

The suggested study outperforms the current proposed study in various areas, providing a comprehensive understanding of technological advancements and research challenges through web mining and heterogeneous online sources. Our system offers real-time analysis, detecting and analyzing new developments promptly, and informing stakeholders.

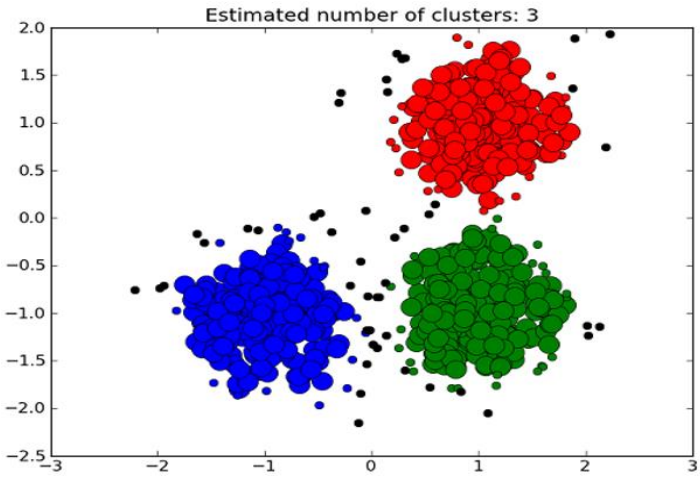

Figure 6: How to Cluster the Data Mining Data

Conclusion & Future Direction

Web mining is a process of discriminating relevant information accessible from the website. It is the branch of data mining which finds the optimal knowledge patterns from enormous, large data reflected from

web data. Mostly the outliers are created inside the dataset; therefore, the pre-processing and cleaning phases are in practice for carrying the data. The mechanism of web mining performs different important stages where they perform a specific task to solve a data mining challenge. Now, the search goes through a hyper mediated source and is thus derived from the missive knowledge, patterns for access, and usage mining of browser logs, user sessions, and cookies usage. Today, web mining techniques become advanced and improved gradually. The valuable data may be retrieved from web data. It is a very powerful technique. Web mining framework has produced the possibility for users by creating the indicators for websites' data innovative ecosystems. With the development of the innovation framework and the reduction of timeliness, the web base will be able to overcome its limitations.

References

Chen, X., Xie, H., Tao, X., Wang, F. L., & Cao, J. (2024). Leveraging text mining and analytic hierarchy process for the automatic evaluation of online courses. International Journal of Machine Learning and Cybernetics, 1-26.

Du, H., Xing, W., & Zhu, G. (2023). Mining teacher informal online learning networks: Insights from massive educational chat tweets. Journal of Educational Computing Research, 61(1), 127-150.

Das, T., & Kondamudi, S. (2023). Customer Relationship Management in a Data-Driven World: Leveraging Data Mining Tools. Data Science and Intelligent Computing Techniques, 227-233.

Govers, R. (2023). A Validation Approach for Agent-Based Simulation Models Using Process Mining and Data Mining Techniques (Master's thesis, University of Twente).

Gheisari, M., Hamidpour, H., Liu, Y., Saedi, P., Raza, A., Jalili, A., ... & Amin, R. (2023). Data mining techniques for web mining: a survey. In Artificial intelligence and applications (Vol. 1, No. 1, pp. 3-10).

Hegade, P., Joshi, R. M., Hegde, V. G., Kale, T., & Basavaraddi, S. (2021). Po-Miner: A Web Mining Poem Generator and its Security Model. SN Computer Science, 2(5), 401.

Ibrahim, K. K., & Obaid, A. J. (2021, May). Web mining techniques and technologies: A landscape view. In Journal of Physics: Conference Series (Vol. 1879, No. 3, p. 032125). IOP Publishing.

Kayser, V., & Shala, E. (2020). Scenario development using web mining for outlining technology futures. Technological forecasting and social change, 156, 120086.

Kayser, V., & Blind, K. (2017). Extending the knowledge base of foresight: The contribution of text mining. Technological Forecasting and Social Change, 116, 208-215.

Kumar, M. R., Venkatesh, J., & Rahman, A. M. Z. (2021). Data mining and machine learning in retail business: developing efficiencies for better customer retention. Journal of Ambient Intelligence and Humanized Computing, 1-13.

Kumar, S., & Kumar, R. (2021). A study on different aspects of web mining and research issues. In IOP conference series: materials science and engineering (Vol. 1022, No. 1, p. 012018). IOP Publishing.

Kumar, R., & Sharma, M. (2016). Advanced neuro-fuzzy approach for social media mining methods using cloud. International Journal of Computer Applications, 975, 8887.

Lee, M., Kim, S., Kim, H., & Lee, J. (2022). Technology opportunity discovery using deep learning-based text mining and a knowledge graph. Technological Forecasting and Social Change, 180, 121718.

Mele, I., Bahrainian, S. A., & Crestani, F. (2019). Event mining and timeliness analysis from heterogeneous news streams. Information Processing & Management, 56(3), 969-993.

Mohammadi, E., & Karami, A. (2022). Exploring research trends in big data across disciplines: A text mining analysis. Journal of Information Science, 48(1), 44-56.

Murtaza, S., & Ahmed, S. (2020). Impact of the Semantic Web mining by using different techniques-A Survey. International Journal of Science and Innovative Research, 1(02).

Rhayem, A., Mhiri, M. B. A., & Gargouri, F. (2020). Semantic web technologies for the internet of things: Systematic literature review. Internet of Things, 11, 100206.

Schedlbauer, J., Raptis, G., & Ludwig, B. (2021). Medical informatics labor market analysis using web crawling, web scraping, and text mining. International Journal of Medical Informatics, 150, 104453.

Shi, F., Chen, L., Han, J., & Childs, P. (2017). A data-driven text mining and semantic network analysis for design information retrieval. Journal of Mechanical Design, 139(11), 111402.

Velkumar, K., & Thendral, P. (2020, March). A survey on web mining techniques. In 2nd international conference on new scientific creations in engineering and technology (ICNSCET-20) International Journal of Recent Trends in Engineering & Research (IJRTER). Special Issue (pp. 2455-1457).