Journal of Computer Science

Review Article

Web Scraping for Product Recommendations: A Review of Techniques and Applications

Amarinder Kaur and Deepak Prashar

School of Computer Science and Engineering, Lovely Professional University, Phagwara, India

Article history
Received: 24-10-2024
Revised: 15-01-2025
Accepted: 06-02-2025

Corresponding Author:
Amarinder Kaur
School of Computer Science and Engineering, Lovely Professional University, Phagwara, India
Email: amarinder.21482@lpu.co.in

Abstract: A comprehensive method for developing a reliable product recommender system leveraging sophisticated web scraping technologies, machine learning, and natural language processing techniques. The proposed system addresses key challenges in personalized product recommendation, including the difficulty of integrating diverse data from multiple e-commerce platforms, ensuring data quality, and improving recommendation accuracy to enhance user experience. Specifically, this research tackles issues related to the heterogeneous nature of data sources, the need for accurate sentiment analysis from textual reviews, and the necessity for dynamic, adaptive recommendation mechanisms that respond to evolving user preferences. The structural setup of the method is composed of three primary stages. In the first stage, data collection from various e-commerce platforms is performed within the limits of legal and ethical guidelines. Tools like Beautiful Soup, Scrapy and Selenium are used to gather comprehensive product data, including descriptions, user reviews, ratings, and metadata. This data undergoes intensive cleaning and preprocessing to ensure high-quality inputs for subsequent stages. Pre-training exploratory data analysis utilizes visualization tools such as Matplotlib and Seaborn to uncover patterns and insights from the data. In the second stage, machine-learning techniques are applied to build effective recommendation models. A collaborative filtering approach, using matrix factorization, predicts interactions between users and items based on historical trends. Concurrently, a content-based filtering approach employs cosine similarity to identify similar items. Additionally, a Natural Language Processing (NLP) approach conducts sentiment analysis, incorporating TF-IDF recommendation algorithms to capture textual preferences from user reviews. Model training and optimization, utilizing frameworks like Tensor Flow or PyTorch, refine the models to maximize applicability and relevance. Evaluation metrics such as MAE and RMSE validate the model performance against known benchmarks, ensuring accurate and personalized recommendations. The third stage emphasizes continuous improvement by creating a feedback loop from user interactions, enabling adaptive preference learning. This adaptive mechanism leverages reinforcement-learning techniques to refine recommendations dynamically based on evolving user behavior and market trends. Ethical considerations are integral to this research, focusing on data transparency, privacy, and adherence to guidelines. This structured and systematic methodology not only the development of personalized recommendation systems but also paves the way for innovations in deep learning for pattern recognition and reinforcement learning for adaptive decision-making. Consequently, this research contributes to the global landscape of e-commerce by enhancing user satisfaction and optimizing product discovery.

Keywords: Recommender Systems, Web Scraping, Machine Learning, Natural Language Processing (NLP), E-commerce

DOI: 10.3844/jcssp.2025.1425.1439

Introduction

Recommender systems are the cornerstone of a flourishing e-commerce ecosystem and are essential for user satisfaction and profit. Unsurprisingly, these systems require copious amounts of user data to offer situational-specific product recommendations that they find the most useful for each unique customer experience on their platform. Although collaborative or content-based methods serve as a classic recommendation system base, many challenges remain unaddressed, such as the cold start issue and lack of diversity in collections. In this study, we investigate a novel product recommendation system utilizing web scraping advancements blended with cutting-edge algorithmic concepts from machine learning (Lee *et al.*, 2022). They serve two main purposes: First, to survey Several Recommender Systems (RSs) in terms of their methodology and pros and cons; the second target is replicating an RS, but this time a more value-added one by integrating collaborative filtering, content-based filtering and NLP for recommendations. The proposed system aims to enhance the accuracy and relevancy of recommendations by utilizing user reviews extracted through web scraping. We tackle challenges such as maintaining data integrity during web scraping, solving the cold start problem with recommendation strategies and crafting an efficient way to scale large-scale data processing by optimizing our computational resources (Fikri *et al.*, 2022). This research seeks to empirically demonstrate their feasibility and validity in e-commerce settings with comprehensive testing and by outperforming benchmark systems, targeting different user profiles, preferences and behaviours.

The major problem with the conventional web scraping technique, which was the main way to get data from online sources in the past, is greatly hindered due to many limitations present while ingesting a website. Usually, such methods are nothing but handling a process that extracts specific data from specified websites using some basic level of HTML parsing, which might not be helpful for scraping modern websites these days. The only limitation is that they cannot deal with interactions in JavaScript (and, let's be real, most of the web uses JS). The problem with traditional scraping methods is that they often cannot grab all the necessary information, especially if a website uses JavaScript to load dynamic content, leaving us with paper-thin results and inaccurate data. On top of that, the old scraping way does not scale very well because web page layouts change all the time (Tabaku & Ali, 2021). We also need to keep in mind that websites are regularly altered and these updates make life difficult for static scraping algorithms when it comes to staying up-to-date with proper content extraction. Traditional scraping methods do not adapt well, which lowers scalability and reliability, causing them to struggle to collect real-time data across a wide array of online sources. One of the disadvantages inherent with traditional scrapers is that they are susceptible to measures website owners may use against scraping. With the increase in scraping activities, various countermeasures have been designed by website administrators to spot and prevent automated bots. These measures vary from IP blocking to CAPTCHA challenges and sophisticated bot detection algorithms. Hence, it becomes difficult for the conventional scraper to avoid detection and collect data in the desired manner.

Importance of Product Recommender Systems in E-commerce

Product Recommender Systems play an important role in e-commerce, having a great impact on user experience and business performance. Product Recommender Systems are a key ingredient for e-commerce platforms and have an immense impact on user experience as well as massive ramifications for the business itself (Pawar & Chiplunkar, 2022). The purpose of these systems is to understand user behaviour and choices, then use that information to provide relevant products to keep engagement high and satisfaction among users active and hence wanting to purchase something, contributing to sales. They are utilized to enhance the customer experience and conversion rates for businesses through tailor-made shopping experiences.

Significance of Using Web Scraping Techniques for Data Collection

Sourcing vast amounts of real-time data from different e-commerce stores is made possible by web scraping. Web scraping enables the capturing of large sets of real-time data from various e-commerce stores. Companies and researchers can use this to compile massive datasets containing product descriptions, user reviews/ratings, metadata, etc., (Nurkholis *et al.*, 2023). This type of information helps in a nuanced knowledge about user intermediate preferences and, ultimately, market trend data on which you model your end recommendation system ads. It also helps in getting real-time data as opposed to old records.

Challenges and Solutions

Implementing a product recommender system involves various challenges stemming from the use of web scraping technologies. These include data quality and integrity issues, the cold start problem, scalability concerns and complexities in sentiment extraction and context sensing from user feedback. Websites often differ in structure and employ dynamic content, leading to noisy and inconsistent data that requires extensive preprocessing to ensure standardization. Without proper handling, traditional systems may provide poor

DOI: 10.3844/jcssp.2025.1425.1439

recommendations, especially for new users or items with limited interaction data.

To address these issues, robust data cleaning and pretreatment techniques are essential to resolve discrepancies and manage missing values. Exploratory Data Analysis (EDA) plays a crucial role in understanding data features and informing model decisions. Hybrid recommender systems, combining collaborative filtering with content-based methods or integrating Natural Language Processing (NLP) for user reviews, mitigate the cold start problem and improve recommendation accuracy. Leveraging cloud-based platforms like Google Colab or Jupyter Notebook enhances scalability by enabling efficient real-time data processing. NLP techniques, such as sentiment analysis and topic modelling, help extract relevant features from user feedback, enriching personalization and contextual relevance in recommendations.

Limitations of Traditional Web Scraping Methods

Traditional web scraping methods, once vital for data collection, have become increasingly inadequate in today's dynamic web environment. Early techniques relied on simple scripts and static HTML parsing, which are ineffective against modern websites that use JavaScript-rich interactions and dynamically generated content. Consequently, static scraping algorithms often fail to capture accurate or complete data. Frequent changes in website structures further diminish the reliability of traditional scrapers, requiring constant script updates. These methods also lack scalability and robustness, making them unsuitable for handling large-scale, real-time data extraction. Additionally, traditional scrapers are highly susceptible to anti-scraping measures, including IP restrictions, CAPTCHA challenges and sophisticated bot detection algorithms, which significantly hinder their effectiveness. Modern solutions must prioritize adaptability and intelligence. Techniques like headless browsing with tools such as Selenium or Puppeteer, combined with AI-based detection of structural patterns, enable more resilient scraping systems. Employing proxy management, CAPTCHA-solving services and adaptive algorithms enhances the reliability of data extraction in the evolving digital landscape. By embracing these approaches, scraping systems can overcome the limitations of traditional methods and ensure efficient, scalable data acquisition.

Figure (1) illustrates the capabilities of a platform designed to convert any website or application into a ready-to-use data API, showcasing several key features. The platform employs AI-powered web and app data extraction, leveraging artificial intelligence to ensure accurate and efficient data retrieval from dynamic sources. It also offers database integration capability, allowing seamless storage and management of extracted data within various databases. Users benefit from real-time or near-real-time data access, ensuring that the most current information is available for decision-making. Additionally, the platform supports workflow automation, streamlining data extraction and integration processes to enhance operational efficiency. Operating within legal boundaries, it ensures compliance with data privacy and security regulations, safeguarding sensitive information. Moreover, the platform enables the creation of custom data scrapers, providing flexibility for extracting specific data tailored to unique business needs. These features collectively enhance the platform's adaptability and utility for modern data-driven applications.

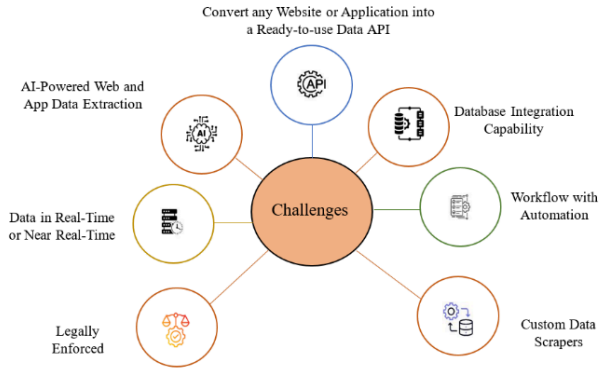

Fig. 1: Challenges in traditional web scraping

Web Scraping Overview

Web scraping, also called webpage extraction, involves gathering data from the internet (often unstructured or semi-structured) and transforming it into a more structured format that can be used for analysis, visualizations, or other purposes. Web scraping allows researchers, developers and companies to gather unstructured data from the web for various purposes, including content aggregation, competitive analysis, market research and machine learning model building. It is very important to respect the rights of website owners and follow moral and legal rules when using web scraping.

Fig. 2: Flow diagram for web scraping

Web scraping, as illustrated in Figure (2), involves several key steps to automate data collection from websites. It begins with requesting web pages, where

HTTP queries are sent to target URLs. Tools like Selenium or programming languages such as Python are commonly used to mimic a browser's behaviour and retrieve webpage content. Once the page content is acquired, HTML parsing follows, extracting relevant information from the HTML markup using libraries like Beautiful Soup and LXML. These tools enable developers to search and access specific elements within the page structure. Next, the parsed data undergoes data extraction, where key items such as text, images, links, or tables are retrieved based on their HTML tags, classes, or IDs. This step can gather diverse information, including product details, prices, reviews and news headlines. After extraction, cleaning and structuring the data is crucial to remove unnecessary elements, handle missing values, standardize formats and organize the data into structured formats like JSON, CSV, or databases. Handling pagination and dynamic content is often necessary for websites using AJAX or dynamic loading techniques. Scraping tools replicate user interactions to capture such content. Finally, scrapers must respect terms of service and robots.txt guidelines, which specify permissible crawling activities, ensuring ethical and legal scraping practices.

Literature Survey

The literature survey explores various contributions and challenges in recommendation systems across different domains. Table (1) shows the key findings that includes the use of graph convolutional networks for group recommendations, hybrid approaches combining knowledge graphs and dynamic interest modelling for restaurant suggestions and trends in automatic recommendation systems for dashboard building. It highlights techniques to improve recommendation accuracy and efficiency, such as filtering data sets and identifying frequent item sets. Additionally, it covers content-based filtering for anime recommendations, sentiment analysis in hybrid movie recommendation systems and the integration of natural language processing for technological recommender systems.

Few analyses of rich data available from online retail Guyt *et al.* (2024) provided a comprehensive review of the relevant literature in their paper. They review the broad applications and uses of retail data through web-scale mining, discuss the hierarchy of e-commerce datasets with respect to detailed web reviews and propose a tutorial methodology for researchers to efficiently harness web-scale digital consumer internet raw materials extracted from an e-shop clone. This study is intended to help those in retail research by providing practical guidance and tools on the use of web data and debunking myths surrounding its application. This is the challenge that preparing master data from existing big data with various formats and on different sites becomes too complicated.

Kudo *et al.* (2022) suggest a methodology of scraping the data needed to normalize, centralize and push it into a semi-structured common database while also extracting individual master data. To guide new learners to Bootstrap, they collected 3238 English questions from a question master for e-learning developed under compilation study and compared MongoDB with MySQL as evidence that semi-structured databases are more efficient when presenting various data.

Rodrigues *et al.* (2024) looks at how we can define perfumes according to their fragrance profile, as well as create new types of odorous molecules that are directly targeted towards consumer preferences. Their method for identifying similar fragrances and helping companies create new ones more likely to meet public preference is constructed by using web scraping for building a perfume dataset, applying k-means within that information and leveraging GNNs, primarily aimed at readers with an interest in Graphs Neural Networks derived from SMILES presentation based on consumer feedback.

The survey conducted by Talari *et al.* (2022) emphasizes the architecture of Big Data and web-based technologies for food safety, with possible effects on climate change factors together with challenges encountered in detecting sources/samples and future research aspects. The foremost challenge is how scattered information, heterogeneous data and complex interactions occur within environmental factors, pollutants and diseases. Insights from this review may assist the agri-food research community in adopting potentially useful Big Data and web-based DSS features for more efficacious food safety risk assessment, with consideration of climate change impacts.

Ghoul *et al.* (2024) emphasize the need to build a sustainable and resilient perfumery-cosmetics supply chain to remain competitive and anticipate market trends. They prototype a semi-automatic AI-based Technological Watch Information System for Surveillance, including defining needs, collecting data, ranking documents and disseminating results. The result is a less costly, automated methodology to discover new ideas and make better decisions.

Londhe *et al.* (2024) specialize in personal travel planning and address this issue directly in their paper 'Personalized Travel Planning' using ChatGPT to understand user preferences and the Playwright framework for real-time data scraping from Google Travel. They train a content-based recommendation engine to map travel options based on user budget and priorities (e.g., pricing, hotel ratings, etc.). The result is a travel planning system achieving 86% recommendation precision, leading to improved user satisfaction through personalized suggestions.

Table 1: Summary of major contribution of the existing studies

Researchers	Contributions	Scope	Advantage	Weakness
Miao *et al*. (2023)	In order to improve restaurant recommendations, this study presents a hybrid recommendation strategy that combines knowledge graphs and dynamic interest modelling	Its goal is to enhance restaurant suggestions by taking into account both dynamic shifts in user interests and restaurant attributes	Extensive experiments demonstrated the effectiveness of the proposed method over baseline approaches	The integration of several cutting-edge approaches can necessitate a sophisticated implementation and significant computer resources
Soni *et al*. (2024)	The research analyzes 19 important papers to find trends and strategies in automatic or semi-automatic recommendation systems for dashboard building.	It focuses on systems that automatically suggest data, optimize layout, integrate user comments and provide visuals for dashboards	outlines potential research possibilities in dashboard recommendation systems and offers a thorough overview of the approaches used today	restricted by the examination of only 19 publications from a starting set of more than 1000, which might not accurately reflect all developments in the subject
Kang and Wang (2024)	By detecting frequently occurring commodity sets and reducing the commodity data set, the suggested technique increases both recommendation accuracy and time efficiency	By filtering data sets, determining user-commodity interest rankings and creating similar product suggestion criteria, the system tackles issues with fusion recommendation methods	Lowering the quantity of candidate frequent item sets makes recommendations that are quicker, more precise and more sensitive to the changing preferences of the user	When examining intricate relationships between customers and items, the algorithm can still run into problems
Marti *et al*. (2023)	The study uses an iterative, transdisciplinary approach to construct efficient earthquake scenarios and quick impact assessments	It caters to the requirements of various target audiences, such as the general public and professional stakeholders	All user groups show that it has a high perceived value and comprehension	It suggests that cartographic data has to be improved and visualizations made simpler. 4o
Reynaldi and Istiono (2023)	The study shows that an anime recommendation system may be successfully merged with a content-based filtering mechanism, as seen by the 74.23% user satisfaction percentage	The study is restricted to assessing user satisfaction with a content-based filtering anime recommendation system employing the Delone and McLean models	The 74.23% satisfaction rate indicates that the system has successfully provided appropriate anime recommendations, suggesting a pleasant user experience	The fact that the study's scope is limited to a certain user group and period of time may reduce how broadly applicable the results can be
Pavitha *et al*. (2022)	In order to improve user experience, the study develops a hybrid movie recommendation system that combines machine learning-based sentiment analysis with Cosine Similarity for suggestions	It compares Naïve Bayes and SVM classifiers and incorporates sentiment analysis to improve movie suggestions	By examining movie reviews, the algorithm provides a recommendation that is better educated, more accurate and more pertinent	The method may be constrained by the sentiment analysis's representativeness and quality as well as the possible computational expense of employing SVM
Campos Macias *et al*. (2022)	The study shows that natural language processing and web crawling can be used to create a technological recommender system	The purpose of the study is to assess the efficacy and efficiency of a prototype system for gathering data from websites pertaining to technology and summarizing it	The prototype effectively extracts text summaries from web pages that contain pertinent search phrases	The results are inadequate due to the random arrangement of extracted sentences in the summary
Vijayakumar and Jagatheeshkumar (2024)	In order to improve individualized learning, the paper presents an e-content recommendation system that divides documents into tiers according to user learning capacity	It uses sophisticated text processing techniques to suggest documents divided into basic, intermediate and levels with the goal of enhancing online learning	When compared to alternative approaches, the system's high accuracy rate of 98% considerably improves the relevance of recommended information	The reliance on SVM and predefined document classifications may hinder adaptation to various and changing learning requirements
Liu *et al*. (2024)	In order to increase recommendation performance, the research presents the GRU-KSC algorithm, which combines an upgraded GRU model with superior spectral clustering	The technique aims to solve cold-start problems in recommendation systems and data sparsity	Combining sophisticated clustering and GRU algorithms enhances recommendation accuracy and robustness	The suggested algorithm's success is dependent on the quality and diversity of the dataset utilized in tests

The study by Rejeb *et al.* (2024) characterizes the issue of ChatGPT's impact on education from a sentiment analysis viewpoint. They processed 2003 web articles using web mining and NLP techniques to analyze opinions. The outcomes illustrate ways in which ChatGPT may aid writing improvement, promote interactive learning and address academic integrity and ethical AI practice. Abdulrahman Rejeb's study provides key insights necessary for best incorporating ChatGPT into the field of education. Currently, the PSS integration with equipment portrait for industrial maintenance is at a low level of research, leading to a lack of refined OM solutions for complex products.

Ren *et al.* (2023) propose a Personalized OM Approach for Complex Products (POMA-CP) that includes a multi-level case library, dynamic equipment portrait model and case-pushing mechanism. The method results in higher accuracy in service schemes, competent OM knowledge reuse and reduced maintenance costs and resource overhead. Shan Ren *et al.*'s work testifies to the effectiveness of this approach. Diverse linguistic characteristics pose challenges in effective multilingual opinion mining from social media data, represented as Marathi and Hindi-based text written using English alphabets.

Shahade *et al.* (2023) propose Collaborative Fine-tuned with Adam (CFA), a novel method combining web scraping, zero-shot instance-weighting and Naïve Bayes vectorization for accurate polarity classification in a deep learning style. The experimental results reveal a high accuracy rate, precision level recall value and F1-Score for the HFS-AO approach, significantly improving computational time compared to PGM, MCM, CNN and NBi-LSTM. This study proposes a four-stage approach to identify the effect of news sentiment on stock prices. It uses web scraping for data extraction, the modified VADER algorithm for sentiment analysis and associational causal analysis to find cause-effect relationships.

The research by Varghese and Mohan (2023) shows that the Renyi entropy approach detects causality and information flow better than the Granger test for stock pairs with fewer media mentions. The model was also applied to the pharmaceutical sector during COVID-19.

Gaffey *et al.* (2023) provide an overview of Green Biorefinery processes, which are intended to create products of a higher quality than current biotechnology methods for sustainable food, feed, or materials in Europe. They explore different models and product opportunities. Green Biorefinery systems demonstrate high potential for sustainable production and new business models but must comply with quality control to enter the market.

Putrama and Martinek (2023) address the challenge of integrating information from educational and job platforms (in free text, across a variety of platforms) into an analysis that will recommend appropriate courses/jobs. They conducted a study using a graph representation learning embedding algorithm to unify the data from these platforms. The method could deliver RMSE scores smaller than 0.1 and AUC scores within 70-75%, highlighting the potential for effective recommendation systems. Managing Unstructured Data (UD) processing is complicated; the process of collection and selection has no clear guidelines. A step-by-step process is proposed for issue identification, message content and message resolution. UD applies to the solutions development framework, encompassing exploration or exploitation and internal or external scanning.

De Haan *et al.* (2024) propose a structured journey guided by this roadmap for utilizing UD as an instrument of choice for making decisions in organizations and enhancing managerial effectiveness with future implications on research. Inefficient water use in irrigation is the result of both limited means and poor practice.

Flores Cayuela *et al.* (2022) created a Decision Support System for Precision Irrigation Management (DSSPIM) based on Information and Communication Technologies, with data provided in real-time from water meters and soil sensors along with methodologies of calculation applied to estimate the use of green water. DSSPIM allowed farmers to identify inefficiencies; it increased field-scale Water Use Efficiency (WUE) and saved 20% of irrigation water when applying orange trees compared to traditional practices. While food recommender systems may provide personal recommendations based on individual preferences, collective needs are complex as they go beyond only personalized taste and pragmatic factors such as health objectives.

Rostami *et al.* (2024) propose that the similarity and preferences of social community users are analyzed using a feature learning and neural network-based model. The novelty of their objective is the introduction of a rate prediction that balances group preferences with health factors. Abolghasemi *et al.* (2024) identify the limitations of recommender systems, such as scalability, cold-start and sparsity, which complicate the selection of techniques for application-focused systems. The paper systematically reviews recent contributions, analyzing applications, algorithms, datasets and performance metrics to frame a taxonomy for effective recommender systems. The review highlights the current research state, identifies gaps and provides insights for developing more efficient recommender systems.

Roy and Dutta (2022) identify the challenge of creating an inclusive and accessible multilingual e-learning environment for universities. The study proposes a web crawling and scraping-driven method for

constructing a multilingual ontology tailored for university e-learning. The approach enhances university e-learning by enabling a globally accessible, continuously updated knowledge repository, transcending linguistic and cultural barriers.

Barwary *et al.* (2023) identify the challenge of optimizing e-commerce product recommendations using social network data from Twitter. The study integrates a directed multilayer network analysis with an e-commerce recommender system, modelling user interactions and conversations on Twitter to deliver socially informed product recommendations. The approach enhances e-commerce recommendations by leveraging social network insights and providing personalized, targeted suggestions based on user interactions and purchase history.

Akshay *et al.* (2024) identify the challenge of integrating internet data into traditional official statistics, particularly for analyzing the rapidly changing rental market. The study uses web scraping to gather data from online real estate portals in Berlin, developing a semi-parametric model to predict rental prices per square meter based on apartment features and location. The model reveals significant differences between online rental offers and existing flat contracts, highlighting the potential and challenges of using internet data in official statistics.

Meyberg *et al.* (2024) address the challenge of integrating face recognition with web scraping to enhance biometric security technologies. The study proposes a face recognition model using YOLO and fine-tunes hyperparameters with a custom dataset. The web scraping system is evaluated for precision and performance metrics include mAP and recall. The optimal configuration achieved an mAP of 0.90 and an average precision rate of 0.87, contributing to advancements in biometric security technology.

Methodology

Three primary parts comprise the methodology for creating a strong product recommender system: Web scraping techniques, machine learning models and Natural Language Processing (NLP) techniques. First, data acquisition entails compiling extensive datasets from various e-commerce sites while abiding by legal and ethical guidelines to guarantee adherence to site-specific terms of service (Mahmuddah *et al.*, 2022). Using web scraping tools like Beautiful Soup and Scrapy, this phase entails extracting product metadata, ratings, user reviews and descriptions.

The next steps involve thorough cleaning and preparation techniques designed to improve the quality and dependability of the data by addressing inconsistencies, eliminating duplicates and imputing missing values. Then, using visualization tools like Matplotlib and Seaborn, Exploratory Data Analysis (EDA) approaches are used to find patterns and insights that are essential for the construction of a model later on.

Machine learning techniques are essential for creating efficient recommendation systems throughout the model construction stage. In an effort to uncover hidden variables influencing user preferences, collaborative filtering techniques use matrix factorization techniques to forecast user-item interactions based on past data. At the same time, content-based filtering makes use of user interactions and product attributes to provide tailored recommendations by using cosine similarity calculations to propose things that are similar to those that the user has interacted with.

By identifying complex user preferences and feelings embedded in reviews, Natural Language Processing (NLP) techniques—sentiment analysis using tools like Vader sentiment analyzer and TF-IDF for extracting relevant features from textual data—further improve recommendation accuracy (Hadasik, 2024).

Using potent frameworks like TensorFlow or PyTorch, model training and optimization are carried out, with an emphasis on optimizing algorithms to maximize the relevance and utility of recommendations. To guarantee scalability and performance, the iterative refinement process includes adjusting model parameters and maximizing computational efficiency. Metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are used to rigorously evaluate the recommender system's effectiveness.

The results are compared to those of other systems to confirm the recommender system's superior performance in providing customized recommendations that are suited to the needs of each individual user. User feedback loops are the engine behind the recommender system's constant improvement and adaptation to changing user preferences and market conditions. Throughout the whole development lifecycle, ethical issues are of utmost importance. These include safeguarding data privacy, maintaining transparency in recommendation procedures and adhering to regulatory frameworks (Lotfi *et al.*,2021).

This methodical approach not only lays the groundwork for future improvements by using cutting-edge AI approaches, but it also predicts future improvements in recommender system design, like reinforcement learning for flexible recommendation schemes and deep learning for improved pattern recognition.

This approach is the frontier of digital commerce and consumer engagement paradigms globally by utilizing web scraping, machine learning and natural language processing techniques in a systematic framework. It not only addresses current challenges but also sets the standard for future innovations in e-commerce and personalized recommendation systems.

Core Equations for Recommender Systems

Collaborative Filtering (CF)

Collaborative Filtering using matrix factorization predicts user-item interactions by modelling the underlying structure of preferences in a latent space. The prediction of a user's rating for an item is based on a combination of factors. These include the global average rating, capturing the overall tendency of users to rate items and biases that account for individual user tendencies and item-specific popularity. Each user and item is represented in a latent factor space, characterized by the user and item latent factor vectors. The interaction between these vectors, computed as the dot product, reflects the alignment between a user's preferences and an item's attributes. The model effectively combines these components—global average, user and item biases and the latent factor interaction—to estimate the rating. This approach captures both explicit and implicit patterns in the data, enabling the system to identify nuanced user-item relationships beyond direct interactions, thereby enhancing the accuracy of recommendations.

User-Item Interaction Prediction (Matrix Factorization):

$\hat{r}_{ui} = \mu + b_u + b_i + q_i^T p_u$

where:

- $\hat{r}_{ui}$is the predicted rating of user for item
- μ is the global average rating, 021+
- b_u is the user bias
- b_i is the item bias
- q_i is the latent factor vector for item
- $\mathrm{P_U}$ is the latent factor vector for user

Content-Based Filtering

Content-based filtering employs mathematical techniques to assess the similarity between items based on their features. A common approach involves using cosine similarity, which measures the angle between two feature vectors representing different items in a multidimensional space. By comparing the vectors' directional alignment rather than their magnitude, cosine similarity determines how closely related two items are in terms of their attributes. This technique normalizes the data, ensuring that similarity is based solely on feature relevance, independent of scale. When applied, items with the highest similarity scores to those a user has previously interacted with are recommended, facilitating precise and personalized suggestions.

Cosine Similarity for Item Similarity:

$Sim\left(i,j\right) = \frac{q_i \cdot q_j}{\|q_i\| \|q_j\|}$

where, q_i and q_j are feature vectors representing items and , respectively.

Natural Language Processing (NLP)

Natural Language Processing (NLP) leverages techniques like Term Frequency-Inverse Document Frequency (TF-IDF) to quantify the importance of specific terms within a document relative to a larger corpus. Term Frequency (TF) measures how frequently a term appears in a given document, reflecting its local significance. Inverse Document Frequency (IDF), on the other hand, evaluates how unique or rare a term is across the entire dataset, assigning higher weights to terms that appear less frequently across documents. The product of TF and IDF provides a weighted score that highlights terms that are both locally important and globally distinctive, making TF-IDF a powerful feature extraction method for tasks such as text classification, sentiment analysis and information retrieval.

Term Frequency-Inverse Document Frequency (TF-IDF):

$TF - IDF\left(t,d\right) = TF\left(t,d\right) \times IDF\left(t\right)$

where:

- $TF\left(t,d\right)$ is the term frequency of term in document
- $IDF\left(t\right)$ is the inverse document frequency of term across all documents

Sentiment Analysis

Sentiment analysis involves calculating the sentiment score of a given text by assessing the polarity and intensity of individual words within the text. The process utilizes the Vader sentiment analysis tool, which evaluates the polarity of each word, representing whether it conveys a positive or negative sentiment. Each word’s polarity is then multiplied by its sentiment intensity, known as "valence," which indicates the strength of the sentiment conveyed by the word. The sentiment score for the entire text is derived by summing the contributions of all words, with the sentiment intensity of each word scaled by its polarity. This method allows for a nuanced assessment of the overall sentiment, providing a quantitative measure of the text's emotional tone, whether positive, negative, or neutral. The resulting sentiment score can be used to gauge the general sentiment of the text, enabling deeper insights into user opinions or reviews.

Sentiment score calculation using Vader sentiment analysis:

$Sentiment\ score = \sum_{i=1}^{n} \left(p_i \times valence_i\right)$

where, p_i is the polarity of the i^{th} word and is its sentiment intensity.

Evaluation Metrics

Evaluation metrics play a crucial role in assessing the performance of recommender systems. Two commonly used metrics are Mean Absolute Error (MAE) and Root

Mean Squared Error (RMSE). MAE measures the average magnitude of the errors between predicted ratings and actual ratings, providing a straightforward evaluation of prediction accuracy. It calculates the average of the absolute differences between predicted and actual ratings, offering insight into the overall prediction error without considering the direction of the error. RMSE, on the other hand, emphasizes larger errors by squaring the differences between predicted and actual ratings, which results in a higher penalty for larger discrepancies. It provides a more sensitive measure of prediction accuracy, as it disproportionately penalizes large errors compared to MAE. Both metrics are computed by averaging the errors across all ratings in the dataset, with MAE offering a linear scale of error magnitude and RMSE providing a more comprehensive evaluation by accentuating significant prediction errors. Together, these metrics help in evaluating the quality of the recommendations generated by the system, guiding the refinement and optimization of the model.

Mean Absolute Error (MAE):

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{r}_i - r_i|$$

where:

- $\hat{r}_i$ is the predicted rating
- r_i is the actual rating
- N is the total number of ratings

Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{r}_i - r_i)^2}$$

Recommender Systems

A product recommender system is a sophisticated instrument that analyzes different kinds of data to offer customers customized product recommendations. Its main objective is to improve the user experience by offering products that are appropriate for each person's tastes and habits. This is a thorough explanation of how such a system works and is created.

Collaborative filtering: The interactions between users and items are the foundation of this strategy. It operates under the presumption that two users are likely to have similar tastes in the future if they have comparable preferences or behaviours in the past. Collaborative filtering can be item-based, suggesting products similar to those a user has liked, or user-based, proposing goods liked by similar people. For instance, if User A and User B both give similar product ratings, User B may suggest a new product to User A as a result of User B's positive reviews.

Content-based filtering: This strategy emphasizes user preferences and item qualities. It suggests goods that are comparable to ones that a user has previously interacted with or found appealing. For example, based on item criteria such as category, brand, or features, the system may recommend new devices or related accessories to a customer who often buys electronic gadgets.

Hybrid models: These systems take advantage of the advantages of both content-based and collaborative filtering. Hybrid models combine various recommendation techniques in an effort to deliver recommendations that are more pertinent and accurate. To increase the overall quality of recommendations, this may entail combining algorithms or use one technique to enhance another.

Development Stages

Data collection: The initial phase entails compiling information from several sources. This covers product features like category, price and specs; user activities like clicks, purchases and ratings; and contextual data like location or time of day. User activity logs, online scraping, third-party databases and APIs are some of the places where data can be gathered.

Data preprocessing: The raw data needs to be cleaned and made ready for analysis when it is gathered. This entails dealing with missing values, eliminating duplication and fixing discrepancies. Data preparation sometimes involves converting data into an analysis-ready format, including encoding category variables or normalizing numerical values.

Feature engineering: Important characteristics that affect the recommendations are found and extracted at this point. Choosing pertinent data points for the recommendation algorithms' use is known as feature engineering. Features could include things like product ratings, user demographics and past purchasing trends.

Model training: To create the recommendation model, the preprocessed data is subjected to a number of algorithms. This could entail applying content-based strategies (like TF-IDF or word embeddings for textual features) or collaborative filtering approaches (like matrix factorization or closest neighbor algorithms), or a mix of the two. In order to forecast which products a user is likely to be interested in, the model learns from the data.

Model evaluation: Evaluation criteria including precision, recall, F1 score and mean squared error (MSE) are used to evaluate the effectiveness of the recommendation model. These metrics aid in assessing the model's predictive performance and its accuracy in matching user preferences with appropriate products.

Recommendation generation: The system generates recommendations for users based on the trained model. This is using the model to forecast which products, based on a user's profile and past interactions, are most likely to be of interest to that particular user.

Feedback loop: To improve the model, user comments on recommendations are gathered and

DOI: 10.3844/jcssp.2025.1425.1439

examined. This input may come in the form of clicks, ratings, or other interactions. The model is updated on a regular basis to enhance accuracy and add new data, so the recommendations are current and relevant.

Deployment: Integrating the recommender system into a platform or application that users interact with is the last step. This enables users to communicate with the system and get customized recommendations for products. The effectiveness of the system is maintained and opportunities for improvement are identified through ongoing user participation and system performance monitoring.

Figure (3) depicts recommender systems, which are complex tools designed to give users individualized suggestions based on their interests and behaviours. They are divided into three categories: Content-based filtering, collaborative filtering and hybrid approaches. Content-based filtering recommends products that are similar to those that a user has previously liked by assessing item attributes and correlating them to user preferences; for example, if a user frequently watches comedies, comparable movies are suggested. Collaborative filtering, which can be user-based or item-based, uses the behaviour of other users to create suggestions, either by suggesting items enjoyed by similar users or by proposing goods comparable to those Sabesan *et al.* (2023). The user has already liked it. This method can alternatively be model-based, which uses machine learning models to forecast user preferences, or memory-based, which relies on past user interaction data. Hybrid approaches combine content-based and collaborative filtering techniques to improve recommendation accuracy and diversity. Other techniques used to refine the recommendation process and provide more tailored suggestions based on complex patterns and interactions include association techniques, which identify frequent co-purchases or related items and Bayesian networks, which use probabilistic models to represent relationships between users, items and ratings.

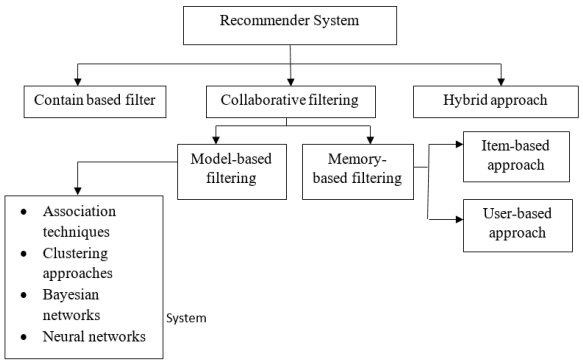

Fig. 3: Recommendation system

Recommender systems are designed to analyze user behaviour and item data to deliver personalized suggestions, enhancing user experience and satisfaction. The three primary types of recommender systems are collaborative filtering, content-based filtering and hybrid models. Collaborative filtering relies on user-item interactions, operating on the premise that users with similar past preferences will likely exhibit comparable future behaviour. It is further categorized into user-based and item-based approaches. User-based filtering identifies similar users and recommends items they have positively interacted with, while item-based filtering suggests items resembling those previously liked by the user. For example, if two users rate similar products highly, one may receive recommendations based on the other's positive experiences. In contrast, content-based filtering focuses on item attributes and user preferences, recommending items akin to those the user previously engaged with. For instance, a user frequently purchasing electronic gadgets may be recommended related accessories based on product features like brand, specifications, or category. Hybrid models combine the strengths of collaborative and content-based filtering to enhance recommendation relevance and accuracy. Hybrid approaches incorporate techniques like matrix factorization or neural networks to address challenges such as the cold start problem, offering more robust and adaptable recommendations across diverse contexts. The key components like collaborative filtering, content-based filtering and NLP techniques are essential for recommender systems but require structural improvements, with clearer transitions and a more logical flow between sections. While mathematical formulations are provided, the technical depth can be enhanced by offering real-world examples and addressing the strengths/limitations of methods. Evaluation metrics like MAE and RMSE need quantitative results and visual representation. Ethical considerations, including data privacy and regulatory compliance, require further discussion. The literature review should integrate recent studies and future directions like reinforcement learning. Including more meaningful figures and diagrams would improve clarity.

Recommendation System Dataflow

When developing a product recommender system utilizing web scraping techniques, the dataflow often begins with data collection, in which web scraping technologies are used to extract relevant information from various online sources such as e-commerce websites, product reviews and social media. This data comprises product descriptions, user reviews, prices and other factors that are required to generate suggestions. Once gathered, raw data is preprocessed to clean and shape it before analysis Park and Shin (2022). This stage entails deleting irrelevant or redundant information, managing missing values and standardizing data formats. Following preprocessing, feature engineering is used to find and choose the most important elements in the data, such as product categories, user ratings and purchase histories (Figure 4). These qualities are critical for

developing good recommendation algorithms. During the model training phase, several algorithms, such as collaborative filtering, content-based filtering and hybrid models, are applied to the prepared data to create a predictive model. This model generates personalized suggestions by learning patterns from prior interactions and item properties. The system's performance is then measured using metrics like precision, recall and F1 score to ensure that the recommendations are correct and relevant. Once validated, the model is utilized to make suggestions to users based on their profiles and interactions. The recommendation system is constantly developed via a feedback loop in which user interactions with the recommendations are tracked and analyzed in order to update and improve the model. The recommender system is connected to a user-facing application, allowing users to receive individualized product recommendations Beveridge *et al.* (2021). Continuous monitoring of user engagement and system performance ensures that the recommendations are current and effective, responding to changing user preferences and market trends.

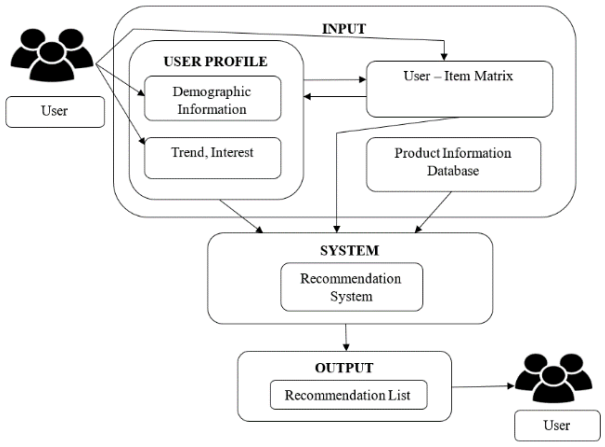

Fig. 4: Recommendation system data flow

Comparative Analysis of Recent Innovations in Recommender Systems

In recent years, the landscape of product recommender systems has seen significant advancements driven by innovative techniques in web scraping, machine learning and Natural Language Processing (NLP). Traditional web scraping methods, such as HTML parsing with BeautifulSoup, have been largely supplanted by modern approaches like headless browsing with Selenium and API-based data extraction, which offer enhanced capabilities for handling dynamic content and complex website structures. In machine learning, collaborative filtering has evolved from matrix factorization methods like Singular Value Decomposition (SVD) to more sophisticated techniques such as neural collaborative filtering, which better capture non-linear user-item interactions. Content-based filtering has similarly progressed, with traditional cosine similarity measures being complemented by neural embeddings, enabling more nuanced item recommendations.

In the realm of NLP, the shift from basic TF-IDF models to transformer-based architectures like BERT has revolutionized sentiment analysis and feature extraction from user reviews, leading to more accurate and context-aware recommendations. Hybrid models, integrating these, are increasingly popular for their ability to address limitations like the cold start problem. These innovations collectively enhance the scalability, accuracy and personalization of recommender systems, setting new benchmarks for performance in the e-commerce domain.

Creating a robust product recommender system integrates web scraping, machine learning and NLP techniques. Data is collected from e-commerce platforms using tools like BeautifulSoup and Selenium, addressing anti-scraping measures (e.g., CAPTCHAs, IP blocking). Preprocessing ensures data quality, while EDA identifies patterns. Collaborative filtering (SVD) and content-based filtering (TF-IDF) form the recommendation backbone, enhanced by hybrid approaches for scalability and cold start problems. NLP techniques (e.g., Vader, BERT) analyze reviews for improved accuracy. Models are optimized using TensorFlow and evaluated with metrics like Precision and NDCG. Ethical practices and future innovations (reinforcement learning, deep learning) are emphasized.

Problem Statement

Data collection challenges: There are ethical and legal issues with web scraping because different websites have different terms of service. Ensuring compliance while gathering detailed information from diverse e-commerce platforms is crucial. Robust scraping strategies are necessary to handle dynamic content, such as changes in product availability and pricing, to ensure timely and accurate data. Additionally, data preprocessing complexity arises from the noise, missing values and discrepancies inherent in raw web-scraped data. Thorough preparation procedures, including data deduplication, outlier identification and imputation, are essential to clean and standardize the data, thereby improving data quality and eliminating biases that could compromise recommendation accuracy.

Cold start problem and scalability: Addressing the cold start problem, where new users or items lack sufficient interaction history for generating accurate recommendations, requires innovative techniques such as data augmentation, hybrid recommendation algorithms, or leveraging contextual and demographic data for immediate customization. Scalability and performance challenges necessitate optimized algorithms and scalable computing infrastructure to manage massive data volumes effectively, particularly in real-time scenarios. This includes utilizing parallel processing techniques and cloud computing resources for efficient data processing and model training. Enhancing user engagement and

feedback mechanisms through active learning, sentiment analysis and continuous model adaptation based on user interactions is essential for improving system accuracy and user satisfaction Barbera *et al.* (2023). Moreover, ensuring ethical and privacy standards in data usage and recommendation practices is critical. Implementing transparent data handling policies, anonymization procedures and compliance with regulatory standards protects user trust and adherence to data protection regulations.

Future Scope

The future scope of product recommender systems includes the integration of AI techniques, enhanced personalization and improved user interaction. Deep learning architectures like Neural Collaborative Filtering (NCF) and transformer-based models, such as BERT, can increase recommendation accuracy by capturing complex patterns in user behaviour and item attributes. Reinforcement learning techniques enable systems to adapt in real time to user interactions, dynamically optimizing recommendations. Enhanced personalization can be achieved through contextual recommendations that consider user location, time of day and device type, as well as multi-modal recommendations that integrate text, graphics and audio using fusion algorithms. AI-driven user interaction can be improved with conversational AI agents that provide tailored interactions and emotion-aware recommendations that respond to user emotions using affective computing techniques. Explainability and transparency are crucial, with interpretable AI models that explain recommendation reasoning and algorithms that mitigate bias to ensure fairness. Unified recommendation systems can offer consistent experiences across various platforms and devices, while cross-platform integration can personalize the customer journey across e-commerce and social media channels Gebretensae (2024). Ethical and regulatory compliance is essential, with a focus on ethical AI practices, user privacy, data protection and adherence to regulations like GDPR and CCPA. Continuous learning and adaptation can be achieved through self-learning systems that automatically adjust to changing trends and feedback-driven optimization that iteratively enhances recommendation algorithms and system performance. The potential developments and advancements that could build upon the current research. It should highlight areas that could enhance the system's performance, functionality and user experience in the future. This includes the integration of cutting-edge AI techniques, such as deep learning models like Neural Collaborative Filtering (NCF) and transformer-based models (e.g., BERT), which could improve recommendation accuracy. Additionally, opportunities for real-time adaptation through reinforcement learning, enhanced personalization through contextual and multi-modal recommendations and AI-driven interactions like conversational agents and emotion-aware systems should be considered. The section should also emphasize the importance of explainability, fairness and compliance with ethical standards, including privacy regulations. Furthermore, cross-platform integration and continuous learning systems should be mentioned as essential for improving user experience and recommendation accuracy over time.

Conclusion

In summary, the implementation of state-of-the-art web scraping techniques and connecting product recommendation engines with different machine learning models alongside Natural Language Processing (NLP) extemporizes user/customer experiences, resulting in more efficient operations within the e-commerce culture. These deliverables detail the methodological rigour and technological innovation inherent in this research approach, which is delineated across three waves of data acquisition and preprocessing, model development and training, as well as evaluation and refinement. On the journey, they have solved challenges like legal and ethical aspects of data acquisition by adhering to compliance practices, navigating through complexity in preprocessing steps with robust cleaning methodologies and solving cold start issues by combining various recommendation strategies. Leveraging scalable cloud computing resources and optimized algorithms to resolve any scalability and performance concerns ensures that the system is able to process large datasets and supports real-time processing requirements. You can simply predict afoot tendencies and look for suggesters by effectively incorporating Deep Learning or Reinforcement teaching machines into one with progressed AI tactics, from multi-modal recommendations to conversational AI and emotion-aware systems that will make personalization not only the only way for users but also more personalized (not just because it happens on mobility through mobile phones) underlining several use cases from content consumption to user satisfaction with a truly delightful experience. Additionally, by maintaining explainability, fairness and regulatory compliance, we maintain trust in the decisions of a recommender system to fulfil ethical standards. Fundamentally, as recommender systems adapt and grow to incorporate the above future directions, they are on a trajectory toward not only meeting user expectations but surpassing them with personalized experiences delivered into richer seams of frictionless digital commerce. By responsibly and considerately innovating into that future, recommendations today will continue to define the consumer experience of tomorrow across global online commerce. The key findings and direct implications of the research make it clear how the current work contributes to the field. For this research, the conclusion should highlight the successful integration of web scraping techniques, machine learning models and NLP in improving product recommendation systems. It should

emphasize the system's ability to enhance user experiences by addressing challenges like ethical data acquisition, preprocessing complexities and cold-start issues. The conclusion should also focus on how the system leverages scalable cloud resources to handle large datasets and real-time processing needs. By incorporating AI techniques, such as deep learning and reinforcement learning, the research demonstrates a strong foundation for personalized and dynamic recommendations. The system's commitment to explainability, fairness and compliance withethical standards ensures trust and reliability. Overall, the research shows that product recommender systems, when implemented correctly, can offer significant improvements in user satisfaction and operational efficiency within e-commerce.

Acknowledgement

The author, Amarinder Kaur, sincerely acknowledges the invaluable guidance and support of Dr. Deepak Prashar whose expertise and mentorship significantly contributed to the successful completion of this research.

Funding Information

The authors have no support or funding to report.

Author's Contributions

Amarinder Kaur: Conducted the literature review, design the methodology, drafting the manuscripts.

Deepak Prashar: Provided guidance on research, direction offered for critical revision, ensured the academic quality and integrity of the work.

Ethics

Author confirms that no ethical issues arise after the publication of this manuscript.

References

Abolghasemi, R., Viedma, E. H., Engelstad, P., Djenouri, Y., & Yazidi, A. (2024). A Graph Neural Approach for Group Recommendation System Based on Pairwise Preferences. *Information Fusion*, *107*, 102343. https://doi.org/10.1016/j.inffus.2024.102343

Akshay, R. K. P., Rinu, R. T. R., Paul, R. T., & Joy, J. (2024). *E-Commerce Recommender System on Twitter using Directed Multilayer Network*. https://doi.org//10.21203/rs.3.rs-4223941/v1

Barbera, G., Araujo, L., & Fernandes, S. (2023). The Value of Web Data Scraping: An Application to TripAdvisor. *Big Data and Cognitive Computing*, *7*(3), 121. https://doi.org/10.3390/bdcc7030121

Barwary, M. J., Jacksi, K., & Al-Zebari, A. (2023). Constructing a Multilingual E-Learning Ontology through Web Crawling and Scraping. *International Journal of Communication Networks and Information Security (IJCNIS)*, *15*(3), 137-153. https://doi.org/10.17762/ijcnis.v15i3.6241

Beveridge, A., Studies, W., & Gallagher, J. (2021). Project-Oriented Web Scraping in Technical Communication Research. *Journal of Business and Technical Communication*, *36*(2), 231-250. https://doi.org/10.1177/10506519211064619

Campos Macias, N., Düggelin, W., Ruf, Y., & Hanne, T. (2022). Building a Technology Recommender System Using Web Crawling and Natural Language Processing Technology. *Algorithms*, *15*(8), 272. https://doi.org/10.3390/a15080272

de Haan, E., Padigar, M., El Kihal, S., Kübler, R., & Wieringa, J. E. (2024). Unstructured Data Research In Business: Toward A Structured Approach. *Journal of Business Research*, *177*, 114655. https://doi.org/10.1016/j.jbusres.2024.114655

Fikri, M. R., Handayanto, R. T., & Irwan, D. (2022). Web Scraping Situs Berita Menggunakan Bahasa Pemograman Python. *Journal of Students' Research in Computer Science*, *3*(1), 123-136. https://doi.org/10.31599/jsrcs.v3i1.1514

Flores Cayuela, C. M., González Perea, R., Camacho Poyato, E., & Montesinos, P. (2022). An Ict-Based Decision Support System for Precision Irrigation Management in Outdoor Orange and Greenhouse Tomato Crops. *Agricultural Water Management*, *269*, 107686. https://doi.org/10.1016/j.agwat.2022.107686

Gaffey, J., Rajauria, G., McMahon, H., Ravindran, R., Dominguez, C., Ambye-Jensen, M., Souza, M. F., Meers, E., Aragonés, M. M., Skunca, D., & Sanders, J. P. M. (2023). Green Biorefinery Systems for The Production Of Climate-Smart Sustainable Products from Grasses, Legumes and Green Crop Residues. *Biotechnology Advances*, *66*, 108168. https://doi.org/10.1016/j.biotechadv.2023.108168

Gebretensae, Y. (2024). Understanding the Cultural Crisis: A Web Scraping Analysis of COVID-19 Vaccine Perceptions and Media Patterns. *Research Square*. https://doi.org/10.21203/rs.3.rs-4297475/v1

Ghoul, D., Patrix, J., Oulmakki, O., & Verny, J. (2024). Information System of Strategic Watch to Rank Innovation Article by Machine Learning Models. *Procedia Computer Science*, *234*, 772-779. https://doi.org/10.1016/j.procs.2024.03.063

Guyt, J. Y., Datta, H., & Boegershausen, J. (2024). Unlocking the Potential of Web Data for Retailing Research. *Journal of Retailing*, *100*(1), 130-147. https://doi.org/10.1016/j.jretai.2024.02.002

Hadasik, B. (2024). *Reduction of Information Asymmetry In E-Commerce: the Web Scraping Approach*.

Kang, L., & Wang, Y. (2024). Efficient and Accurate Personalized Product Recommendations Through Frequent Item Set Mining Fusion Algorithm. *Heliyon*, *10*(3), 25044. https://doi.org/10.1016/j.heliyon.2024.e25044

Kudo, T., Yamamoto, T., & Watanabe, T. (2022). Three-Step Master Data Creation Method from Big Data: Scraping, Semi-Structuring, and Extraction. *Procedia Computer Science*, *207*, 360-369. https://doi.org/10.1016/j.procs.2022.09.070

Lee, M. J., Kang, J., Hreha, K., & Pappadis, M. (2022). A Novel Web Scraping Approach to Identify Stroke Outcome Measures: A Feasibility Study. *Archives of Physical Medicine and Rehabilitation*, *103*(3), 30. https://doi.org/10.1016/j.apmr.2022.01.082

Liu, Q., Yu, M., & Bai, M. (2024). A Study on A Recommendation Algorithm Based on Spectral Clustering and Gru. *IScience*, *27*(2), 108660. https://doi.org/10.1016/j.isci.2023.108660

Londhe, K., Dharmadhikari, N., Zaveri, P., & Sakoglu, U. (2024). Enhanced Travel Experience using Artificial Intelligence: A Data-driven Approach. *Procedia Computer Science*, *235*, 1920-1928. https://doi.org/10.1016/j.procs.2024.04.182

Lotfi, C., Srinivasan, S., Ertz, M., & Latrous, I. (2021). Web Scraping Techniques and Applications: A Literature Review. *SCRS Conference Proceedings on Intelligent Systems*, 381-394. https://doi.org/10.52458/978-93-91842-08-6-38

Mahmuddah, L. A. A., Wibowo, S. A., & Budiman, G. (2022). Generating Information of Url Based on Web Scraping Using Yolov3 Face Recognition Technology. *IJAIT (International Journal of Applied Information Technology)*, *5*(2), 112-122. https://doi.org/10.25124/ijait.v5i02.3910

Marti, M., Dallo, I., Roth, P., Papadopoulos, A. N., & Zaugg, S. (2023). Illustrating the Impact of Earthquakes: Evidence-Based and User-Centered Recommendations on How to Design Earthquake Scenarios and Rapid Impact Assessments. *International Journal of Disaster Risk Reduction*, *90*, 103674. https://doi.org/10.1016/j.ijdrr.2023.103674

Meyberg, C., Rendtel, U., & Leerhoff, H. (2024). Flat Rent Price Prediction in Berlin with Web Scraping. *AStA Wirtschafts- Und Sozialstatistisches Archiv*, *18*(2), 245-278. https://doi.org/10.1007/s11943-024-00340-6

Miao, L., Li, X., Yu, D., Ren, Y., Huang, Y., & Cao, S. (2023). Integrating Users' Long-Term and Short-Term Interests with Knowledge Graph to Improve Restaurant Recommendation. *Journal of King Saud University - Computer and Information Sciences*, *35*(9), 101735. https://doi.org/10.1016/j.jksuci.2023.101735

Nurkholis, A., Fernando, Y., & Ans, F. A. (2023). Metode Vector Space Model Untuk Web Scraping Pada Website Freelance. *INTI Nusa Mandiri*, *18*(1), 52-58. https://doi.org/10.33480/inti.v18i1.4266

Park, Y., & Shin, Y. (2022). Novel Scratch Programming Blocks for Web Scraping. *Electronics*, *11*(16), 2584. https://doi.org/10.3390/electronics11162584

Pavitha, N., Pungliya, V., Raut, A., Bhonsle, R., Purohit, A., Patel, A., & Shashidhar, R. (2022). Movie Recommendation and Sentiment Analysis Using Machine Learning. *Global Transitions Proceedings*, *3*(1), 279-284. https://doi.org/10.1016/j.gltp.2022.03.012

Pawar, S., & Chiplunkar, N. (2022). *Dynamic Searching of Web Services Through Web Scraping*.

Putrama, I. M., & Martinek, P. (2023). Integrating Platforms through Content-Based Graph Representation Learning. *International Journal of Information Management Data Insights*, *3*(2), 100200. https://doi.org/10.1016/j.jjimei.2023.100200

Rejeb, A., Rejeb, K., Appolloni, A., Treiblmaier, H., & Iranmanesh, M. (2024). Exploring The Impact of Chatgpt on Education: A Web Mining And Machine Learning Approach. *The International Journal of Management Education*, *22*(1), 100932. https://doi.org/10.1016/j.ijme.2024.100932

Ren, S., Shi, L., Liu, Y., Cai, W., & Zhang, Y. (2023). A Personalised Operation and Maintenance Approach for Complex Products Based oOn Equipment Portrait Of Product-Service System. *Robotics and Computer-Integrated Manufacturing*, *80*, 102485. https://doi.org/10.1016/j.rcim.2022.102485

Reynaldi, & Istiono, W. (2023). Content-based Filtering and Web Scraping in Website for Recommended Anime. *Asian Journal of Research in Computer Science*, *15*(2), 32-42. https://doi.org/10.9734/ajrcos/2023/v15i2318

Rodrigues, B. C. L., Santana, V. V., Queiroz, L. P., Rebello, C. M., & B. R. Nogueira, I. (2024). Harnessing Graph Neural Networks to Craft Fragrances Based on Consumer Feedback. *Computers and Chemical Engineering*, *185*, 108674. https://doi.org/10.1016/j.compchemeng.2024.108674

Rostami, M., Berahmand, K., Forouzandeh, S., Ahmadian, S., Farrahi, V., & Oussalah, M. (2024). A Novel Healthy Food Recommendation to User Groups Based on a Deep Social Community Detection Approach. *Neurocomputing*, *576*, 127326. https://doi.org/10.1016/j.neucom.2024.127326

Roy, D., & Dutta, Mala. (2022). A Systematic Review and Research Perspective on Recommender Systems. *Journal of Big Data*, *9*(1), 59. https://doi.org/10.1186/s40537-022-00592-5

Sabesan, N., Nivethitha, Shreyah, J. N., Pranauv, A. J., & Shyam, R. (2023). Medical Ministrations through Web Scraping. *ArXiv:2306.12310*. https://doi.org/10.48550/arXiv.2306.12310

Shahade, A. K., Walse, K. H., Thakare, V. M., & Atique, M. (2023). Multi-Lingual Opinion Mining for Social Media Discourses: an Approach Using Deep Learning Based Hybrid Fine-Tuned Smith Algorithm with Adam Optimizer. *International Journal of Information Management Data Insights*, *3*(2), 100182. https://doi.org/10.1016/j.jjimei.2023.100182

Soni, P., de Runz, C., Bouali, F., & Venturini, G. (2024). A Survey on Automatic Dashboard Recommendation Systems. *Visual Informatics*, *8*(1), 67-79. https://doi.org/10.1016/j.visinf.2024.01.002

Tabaku, B., & Ali, M. (2021). Protecting Web Applications from Web Scraping. *Emerging Technologies in Computing*, 56-70. https://doi.org/10.1007/978-3-030-90016-8_4

Talari, G., Cummins, E., McNamara, C., & O'Brien, J. (2022). State of the Art Review Of Big Data and Web-Based Decision Support Systems (Dss) for Food Safety Risk Assessment with Respect to Climate Change. *Trends in Food Science & Technology*, *126*, 192-204. https://doi.org/10.1016/j.tifs.2021.08.032

Varghese, R. R., & Mohan, B. R. (2023). Study on the Sentimental Influence on Indian Stock Price. *Heliyon*, *9*(12), 22788. https://doi.org/10.1016/j.heliyon.2023.e22788

Vijayakumar, P., & Jagatheeshkumar, G. (2024). User's Learning Capability Aware E-Content Recommendation System for Enhanced Learning Experience. *Measurement: Sensors*, *31*, 100947. https://doi.org/10.1016/j.measen.2023.100947