QuickReads: Your Personalized Content Guide using Hybrid Filtering

Sonal Sonarghare
Department of CSE (Data Science),
A.P. Shah Institute of Technology,
Thane (M.H), India 400615
sonalsonarghare432@apsit.edu.in

Sarala Mary
Department of CSE (Data Science),
A.P. Shah Institute of Technology,
Thane (M.H), India 400615
saralas@apsit.edu.in

Sanskruti Chavan
Department of CSE (Data Science),
A.P. Shah Institute of Technology,
Thane (M.H), India 400615
sanskrutichavan443@apsit.edu.in

Janhavi Kasar
Department of CSE (Data Science),
A.P. Shah Institute of Technology,
Thane (M.H), India 400615
janhavikasar429@apsit.edu.in

Harsh Shelke
Department of CSE (Data Science),
A.P. Shah Institute of Technology,
Thane (M.H), India 400615
21107022.harsh.shelke@gmail.com

***Abstract*—This research paper presents the development and evaluation of QuickReads, an innovative hybrid recommendation system tailored to deliver personalized article recommendations. QuickReads employs a synergistic blend of collaborative filtering and content-based filtering techniques, augmented by sophisticated web scraping methodologies to curate a diverse array of pertinent articles from a multitude of sources. The system's robust architecture encompasses meticulous data collection procedures, comprehensive data preprocessing steps, and advanced feature extraction methodologies, all contributing to its robust recommendation engine The system includes features designed to enrich the user experience include summary generation, dictionary lookup, and read-aloud functionality, which collectively enhance usability and engagement. The performance of QuickReads is rigorously evaluated using a diverse set of metrics, including precision, recall, F1-score, and user satisfaction surveys. These evaluations underscore the system's efficacy in providing tailored, compelling content recommendations that resonate with users' preferences and interests.**

Keywords—QuickReads, Hybrid recommendation system, Collaborative filtering, Content-based filtering, Web scraping methodologies, Data collection procedures, Data preprocessing techniques, Feature extraction, User experience, Summary generation, Dictionary lookup, Read-aloud functionality, Performance evaluation, User satisfaction surveys.

I. Introduction

In today's digital age, the vast amount of content available online poses a challenge for users to discover relevant and engaging articles. Recommendation systems play a crucial role in addressing this challenge by providing personalized content suggestions based on user preferences and behaviors. QuickReads emerges as a solution, offering a hybrid recommendation system that combines collaborative filtering and content-based filtering techniques to deliver tailored article recommendations [2].

QuickReads utilizes web scraping techniques to aggregate articles from a wide range of websites, ensuring a comprehensive array of topics and catering to diverse interests. This method not only keeps the system up-to-date with the latest content but also adapts to users' changing preferences over time.

Key features of QuickReads include the dictionary, generation of article summaries and a read-aloud function, enhancing accessibility and usability for users with different preferences and needs. The system's interface is designed to provide a seamless and intuitive browsing experience, allowing users to explore recommended articles effortlessly [1].

Evaluation metrics such as precision, recall, and user satisfaction surveys are employed to assess QuickReads' performance in delivering personalized and engaging content recommendations. The results demonstrate the system's effectiveness in meeting users' information needs and enhancing their browsing experience.

This research paper delves into the architecture, algorithms, evaluation methodologies, and ethical considerations underlying QuickReads. By presenting a comprehensive case study of this hybrid recommendation system, this paper aims to contribute valuable insights to the field of personalized content recommendation and inform future advancements in recommendation system development.

II. Literature Review

The literature review delves into recommendation systems' evolution, vital for personalized content delivery amidst the surging digital content. It encompasses collaborative filtering, content-based filtering, hybrid methods, and advanced techniques like deep learning. This analysis aims to scrutinize key concepts, challenges, and trends while pinpointing areas for future research and enhancements, shaping the understanding and development of recommendation systems.

Essien, Uwah, and Ododo (2021) developed a web-based text-to-speech system for visually impaired users. Using JavaScript, Natural Language Processing, and Digital Signal Processing, the system enables real-time conversion of text to speech, enhancing accessibility across devices. It emphasizes compatibility with speech synthesis libraries and advances in accessibility technology for education and digital environments.

Singhal (2021) presents a research paper on recommendation systems, emphasizing context-awareness and discussing collaborative filtering and content-based approaches. The study reviews existing recommendation methods, analyzes their real-world operation, and highlights the importance of context-awareness. The result underscores a comprehensive understanding of recommendation systems and their pivotal role in information retrieval. [2]

Ye, Tu, and Liang (2019) [3] developed a hybrid recommendation system to address limitations in traditional content-based and collaborative filtering methods for recommending research articles. While content-based filtering relies on item characteristics and collaborative filtering on user behavior, their hybrid system combines both, leveraging the strengths of each approach. The study demonstrates the superiority of their hybrid system over traditional methods, aligning with prior research advocating for hybrid recommendation systems in various domains.

Sen Zhang (2023) [4] delved into enhancing restaurant recommendations using TF-IDF vectorization, merging content-based and collaborative filtering approaches with deep learning. By leveraging TF-IDF to extract textual features from reviews, it recommends restaurants based on similarities in reviews, improving recommendation accuracy and user satisfaction.

The study conducted by Rusdiansyah, Rusdiansyah (2024) explored web program testing using Selenium Python, emphasizing automation for login, transactions, and customer data on website. It discusses best practices, benefits in efficiency, security, user experience, and automated testing, contributing insights for web developers and researchers. [5]

III. SYSTEM ARCHITECTURE

The System Design (refer Fig 1) delineates the step-by-step process, including data collection, preprocessing techniques, feature engineering, model development, and evaluation criteria, all designed to provide a comprehensive understanding of our innovative approach to enhancing QuickReads Hybrid Recommendation System.

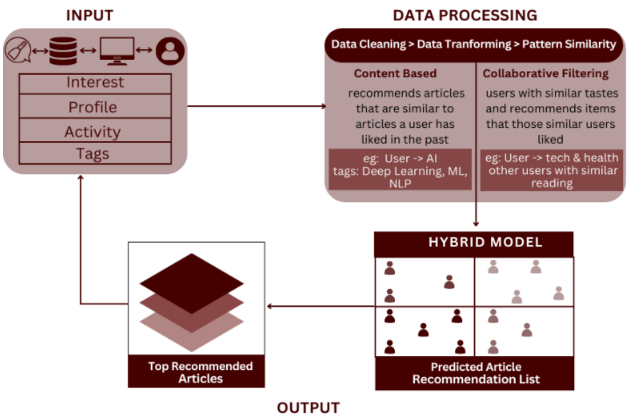


Figure 1. System Design for Hybrid Model

A. *Hybrid Model:*

The architecture (refer Fig 2) illustrates the process of building a hybrid recommendation model for the QuickReads:

1. Data Collection : Article Data is collected using web scraping. For this Selenium tool is used to perform dynamic web scraping of the articles.[5]

2. Data Preprocessing: User interactions data, including user IDs, article IDs, and actions, are collected and preprocessed to remove duplicates and format timestamps. Articles data, such as article IDs, titles, and types are preprocessed to standardize date formats and prepare for feature extraction.

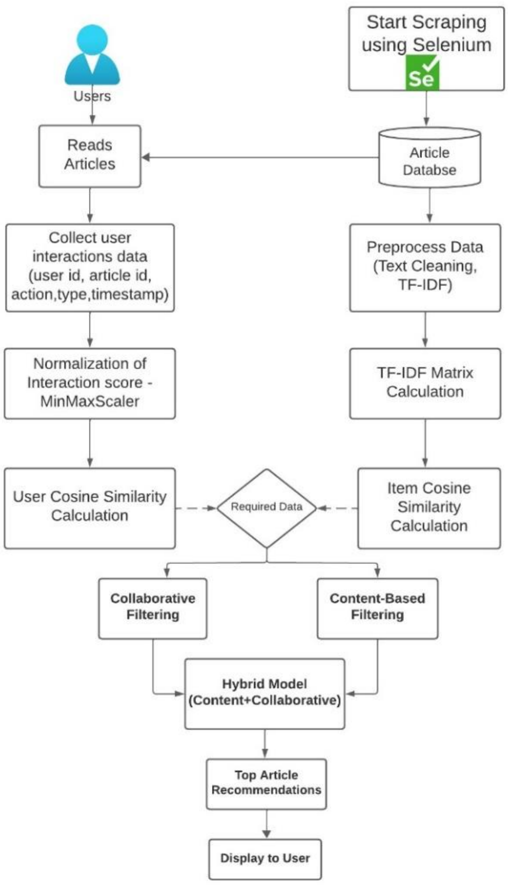


Figure 2. Architecture of the hybrid recommendation system

3. Feature Engineering: TF-IDF vectors are generated for article titles and types using Tfidf Vectorizer. Similarity scores are computed based on article titles using linear_kernels [4]

$$TF(t,d) = \frac{number\ of\ times\ t\ appears\ in\ d}{total\ number\ of\ terms\ in\ d}$$

$$IDF(t) = log\frac{N}{1+df}$$

$$TF-IDF(t,d) = TF(t,d) * IDF(t)$$

where, d refers to a title or type of article,

N is the total number of articles,

df is the number of articles with term t.

4. Normalization and Matrix Creation: Interaction scores are normalized using MinMaxScaler. A user-item interaction matrix is created to represent user-article interactions, and it is converted to a sparse matrix for efficient computation.[3]

5. Cosine-similarity: Model computes similarity scores based on article titles using linear kernel (cosine_sim_title) and article types using linear kernel (cosine_sim_type).

   The cosine similarity formula measures the cosine of the angle between two vectors A and B, representing the similarity between the vectors. In this case, A and B are TF-IDF vectors.

$$\text{cosine\_sim}(A, B) = \frac{A \cdot B}{\|A\| \times \|B\|}$$

6. User Similarity Matrix: User similarity matrix is computed based on interactions using pairwise_distances.

7. Recommendation Functions:
   i. get_collaborative_recommendations: This function utilizes the user_similarity matrix to generate top N recommendations for a user. It calculates the similarity scores between the target user and other users. Identifies the most similar users based on these scores. Aggregates the interaction scores of these similar users with articles to recommend the top N articles that the target user is likely to be interested in.
   ii. get_content_based_recommendations: This function provides top N recommendations for a user using content-based filtering. It identifies articles that the user has viewed. Computes the similarity scores between these viewed articles and all other articles based on their TF-IDF vectors. Recommends articles with the highest similarity scores, excluding those already viewed by the user.
   iii. get_hybrid_recommendation: Combines collaborative and content-based recommendations to provide hybrid recommendations. Removes duplicates and presents the top N hybrid recommendations to the user.[3]

8. Evaluation Metrics: Accuracy metrics such as precision, recall, and F1-score are computed to evaluate the model's performance. Data is split into training and test sets, and recommendations are made for users in the test set to assess the model's accuracy.

IV. RESULT AND DISCUSSION

The experimental results demonstrate the effectiveness of the proposed article recommendation system. The hybrid approach achieves a significant improvement in accuracy compared to individual techniques. The precision, recall, and F1-score metrics indicate the system's ability to make relevant and personalized recommendations to users.

In QuickReads, users have the freedom to choose from various domains or topics of interest (refer Fig 3) right from the start. This user-centric approach enhances the personalization of article recommendations and allows users to tailor their content discovery experience according to their preferences.[3]

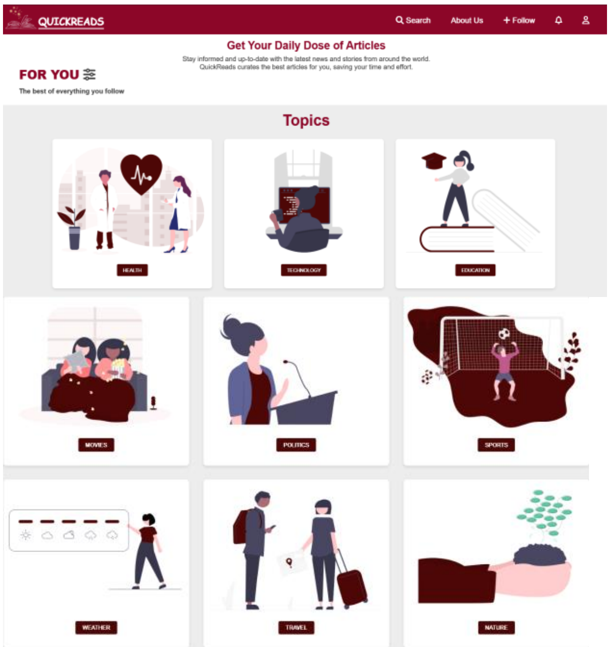


Figure 3. Topics of Interest

QuickReads includes a summary feature [1] to provide users with concise and informative summaries of articles (refer Fig 4), enhancing the browsing experience and facilitating quick comprehension of content. This feature is particularly useful for users who want to grasp the main points of an article without reading the entire text. The process of summarizing text data from a CSV file by utilizing Natural Language Processing (NLP) techniques. It reads text from the 'Content' column of the CSV file, tokenizes the text into words and sentences, removes stop words, and calculates the frequency of words in each sentence. Based on the frequency analysis, it assigns values to sentences and generates a summary by selecting sentences with values above a certain threshold, ensuring that the summary captures the most relevant information. The script then updates the CSV file by adding a new column containing the generated summaries. This approach streamlines the text summarization process, making it efficient for handling large volumes of textual data and extracting key insights from the text for further analysis or presentation.

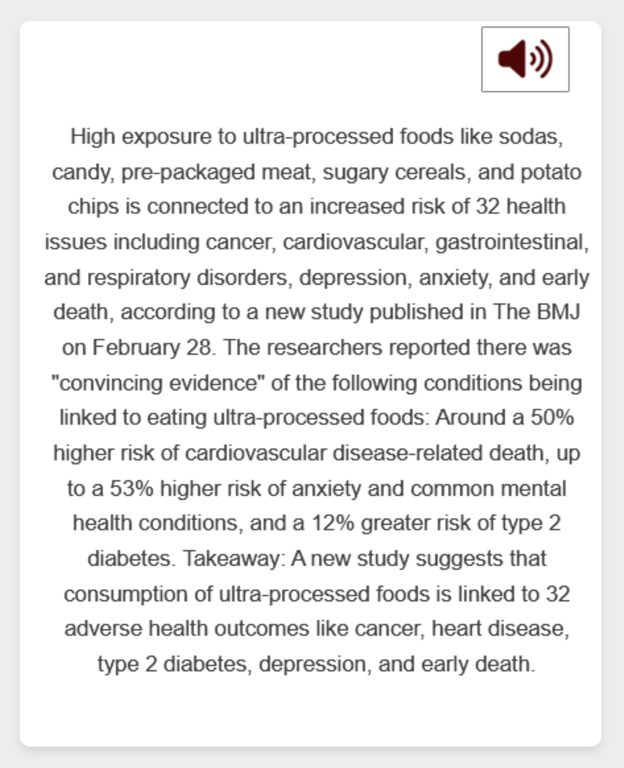


Figure 4. Summary and Read Aloud for Article

The "Read Aloud" option for summaries in QuickReads enhances accessibility and user experience by providing an audio version of the summarized content. This feature is especially beneficial for users who prefer auditory learning styles . When users choose the "Read Aloud" option (refer Fig 4 Read Aloud button) for a summary, QuickReads utilizes SpeechSynthesisUtterance API to convert the summarized text into spoken audio. Users can listen to the summary through their device's speakers or headphones, allowing them to absorb the information hands-free while engaging in other activities.

The dictionary feature allows users to search for word meanings, pronunciation, and definitions using a dictionary API. When a user enters a word in the text box, they receive the following information (refer Fig 5):

1. Meaning: The meaning or definitions of the word.
2. Pronunciation: The pronunciation of the word.
3. Read Aloud: An option to listen to the pronunciation of the word.

This feature enhances the user experience by providing comprehensive information about the word they are searching for, including its definition and how it is pronounced.

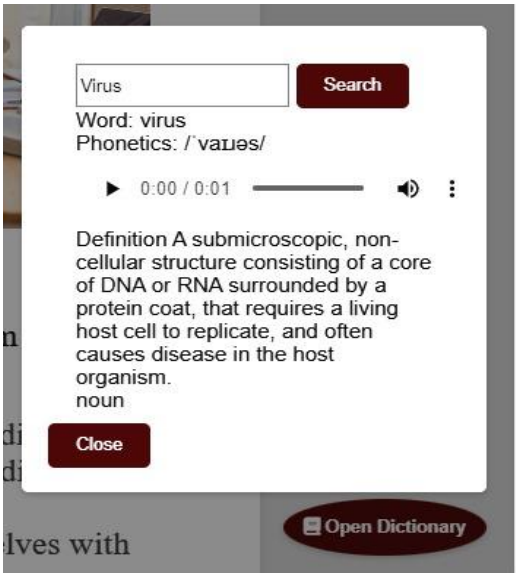
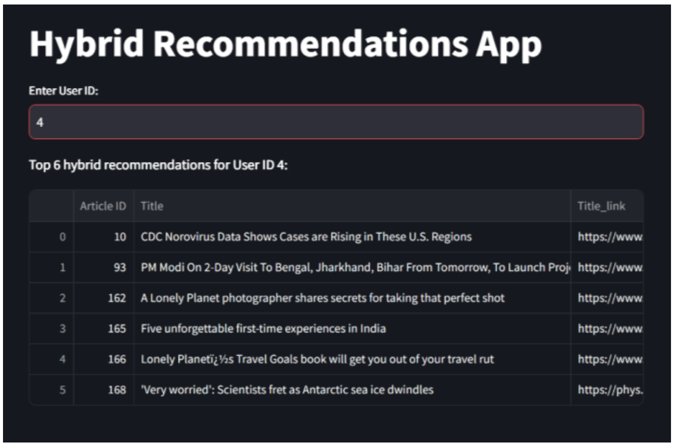


Figure 5. Dictionary

QuickReads gives top 5/6 recommendations to individual users. These recommendations (refer Fig 6) are generated using a hybrid recommendation model, which combines collaborative filtering and content-based filtering techniques to offer more accurate and diverse suggestions. This represents the articles deemed most relevant and interesting for the user based on their past interactions and preferences.



	Article ID	Title	Title_link
0	10	CDC Norovirus Data Shows Cases are Rising in These U.S. Regions	https://www.
1	93	PM Modi On 2-Day Visit To Bengal, Jharkhand, Bihar From Tomorrow, To Launch Proj	https://www.
2	162	A Lonely Planet photographer shares secrets for taking that perfect shot	https://www.
3	165	Five unforgettable first-time experiences in India	https://www.
4	166	Lonely Planetï¿½s Travel Goals book will get you out of your travel rut	https://www.
5	168	'Very worried': Scientists fret as Antarctic sea ice dwindles	https://phys.

Figure 6. Top Recommended article

The accuracy scores offer insights into the system's ability to accurately recommend relevant content to users based on their interactions and preferences. Higher accuracy scores indicate a stronger alignment between recommended items and user interests, contributing to enhanced user satisfaction and engagement.

A histogram (refer Fig 7) is generated to visualize the distribution of accuracy scores among users. This plot provides insights into the spread and concentration of accuracy values, with dashed lines indicating the average and median accuracy values for reference.

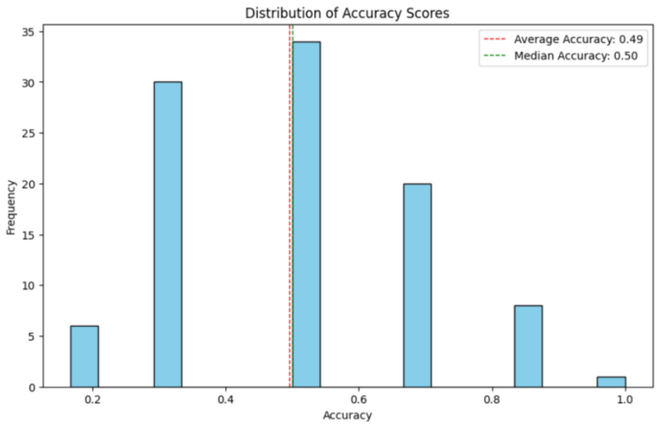


Figure 7. Distribution of Accuracy Scores of Users

The average accuracy score, calculated at approximately 0.49, provides a consolidated measure of the recommendation system's performance across multiple users. This metric represents the system's overall ability to deliver relevant and personalized content recommendations to users based on their interactions and preferences. An average accuracy score of 0.49 indicates that, on average, the system correctly recommends relevant content to users roughly 49% of the time.

Overall, these analyses and visualizations deepens the understanding of the recommendation system's performance

V. CONCLUSION

In conclusion, QuickReads represents an innovative and effective hybrid recommendation system designed to provide personalized article recommendations to users. Through the integration of collaborative filtering, content-based filtering, and web scraping techniques, QuickReads offers a comprehensive solution for content discovery and engagement.

The architecture of QuickReads, including data collection through web scraping, data preprocessing, feature extraction, recommendation algorithms, and user interface design, has been carefully crafted to deliver accurate and relevant recommendations while ensuring a seamless user experience.

Additionally, the system's future scope includes enhancements in personalization, multi-modal content analysis, context-aware recommendations, interactive user feedback, semantic search, generative AI to write articles and ethical AI practices.

Overall, QuickReads stands as a promising recommendation system that not only meets users' information needs but also adapts to evolving technological trends and user expectations. By continuing to innovate and expand its capabilities, QuickReads is poised to remain a valuable tool for personalized content discovery in the digital age.

REFERENCES

[1] Essien, Nesabasi P., Victoria A. Uwah, and Emmanuel P. Ododo. "An interactive intelligent web-based text-to-speech system for the visually impaired." Asia-Africa Journal of Recent Scientific Research 1 (2021): 76-98.

[2] Singhal, A., Rastogi, S., Panchal, N., Chauhan, S., & Varshney, S.(2021). "Research Paper On Recommendation System." Global Scientific Journals.

[3] Ye, Bo Kai, Yu Ju Tony Tu, and Ting Peng Liang. "A hybrid system for personalized content recommendation." Journal of Electronic Commerce Research 20.2 (2019): 91-104.

[4] Zhang, Sen. "Restaurant Recommendation System Based on TF-IDF Vectorization: Integrating Content-Based and Collaborative Filtering Approaches." 2023 International Conference on Data Science, Advanced Algorithm and Intelligent Computing (DAI 2023). Atlantis Press, 2024.

[5] Rusdiansyah, Rusdiansyah, et al. "Web Program Testing Using Selenium Python: Best Practices and Effective Approaches." *Sinkron: jurnal dan penelitian teknik informatika* 8.2 (2024): 994-1000.