A Content-Based Recommender System for News Article Recommendations

Rokkam Nithin Kumar
School of Computer Science and Engineering
Lovely Professional University
Phagwara, India
nithinrokkam@gmail.com

***Abstract*—This paper presents a novel news article recommendation system designed to enhance user experience in content discovery through content-based recommender system. The system leverages a labeled dataset of news articles, employing TF-IDF vectorization and cosine similarity to generate personalized recommendations. By analyzing both the textual content of articles and user interaction data, our approach aims to deliver more relevant and engaging content. Initial evaluations highlight the system's potential to provide meaningful recommendations, aiding users in navigating vast and diverse information landscapes.**

Keywords: News Recommendation System, TF-IDF, Cosine Similarity, User Engagement, Personalized Content

1. Introduction

The rapid expansion of digital news platforms has created an overwhelming volume of content, making it increasingly difficult for users to find articles that align with their interests. Traditional recommendation systems, often based on collaborative filtering, struggle to effectively leverage the detailed information embedded within news articles. This paper introduces a content-based recommendation system that utilizes content analysis and filtering techniques to interpret and utilize the semantic content of news articles directly.

Collaborative filtering methods, which depend heavily on user interaction data, often fall short in the news domain, where relevance is highly dynamic and often depends on current events, trending issues, and emerging themes. These approaches may not capture the nuanced and evolving topics that define news content, as they rely on historical user interactions that may not fully reflect timely shifts in topic relevance. Our proposed system addresses these limitations by focusing on the content within each article, enabling it to provide meaningful recommendations based on the content itself, even in rapidly changing news contexts.

Our approach incorporates vector-based methods to compute content similarity, capturing thematic relationships within articles. By emphasizing the intrinsic content and aligning it with user preferences, our system offers more personalized and relevant recommendations that go beyond simple popularity metrics.

Furthermore, this system adapts dynamically to shifts in user interests, especially in response to trending topics and emerging news themes. Through content analysis, the recommendation engine enhances the relevance and timeliness of suggested articles, creating a personalized experience that feels responsive and engaging. This paper contributes to the field by demonstrating how content-based techniques can improve recommendation quality, presenting a robust alternative to traditional models that may overlook the depth of the articles' content.

2. Related Work

2.1. Traditional Recommendation Approaches

The foundation of recommendation systems typically involves two core methodologies: collaborative filtering and content-based filtering. Collaborative filtering derives patterns from user-item interactions to suggest relevant items (Salakhutdinov, Mnih, Hinton, 2007), with item-based collaborative filtering, such as that used by Amazon, identifying item similarity through co-purchase behavior to scale effectively for large datasets (Linden, Smith, York, 2003; Sarwar et al., 2001). In contrast, content-based filtering utilizes item features to recommend content aligned with user-specific interests, tapping into textual or metadata attributes (Ricci et al., 2015; Van Meteren Van Someren, 2000).

2.2. News Article Recommendation Systems

Several studies have specifically explored recommendation systems for news articles, employing content-based approaches that analyze the unique features of each article. Zhang et al. (2018), for example, leveraged article content attributes to match users with relevant news topics, enhancing user engagement by focusing on content alignment. Wang et al. (2020) proposed models that merge content-based and collaborative filtering elements, though the content-based aspect remains essential for delivering timely, relevant recommendations in news contexts. While hybrid models combine multiple approaches, content-based systems can uniquely cater to the fast-paced, ever-evolving nature of news content by directly analyzing the intrinsic features of articles

2.3. Research Gaps and Study Contributions

While collaborative and hybrid models are commonly implemented in recommendation systems, there is limited research on fully leveraging content-based methods to analyze the intrinsic characteristics of news articles. This study contributes to the field by demonstrating how content-driven techniques can offer a robust alternative, particularly in scenarios where user data is sparse or new topics emerge frequently. By focusing on the content attributes of articles, this approach addresses the unique demands of news recommendation systems, ensuring relevance even in dynamic news environments.

3. Methodology

3.1. Dataset Description

This study utilizes a dataset comprising a comprehensive collection of labeled news articles that spans a broad range of topics and publication sources, as well as information on user interests and activities. The dataset includes articles labeled by topic, title, and associated links, providing a well-organized structure for analyzing relationships between diverse content categories. For this study, the dataset was filtered to include only English-language articles, focusing on topics such as politics, technology, health, and entertainment. This diverse selection ensures that the recommendation system is robust across various content domains, allowing for thorough evaluation and validation.

3.2. System Architecture

3.2.1. Data Preprocessing. The data preprocessing stage involves a series of steps to clean and structure the dataset, ensuring high data quality and relevance for analysis. This includes removing duplicate records, handling missing values, and filtering out low-quality or irrelevant articles. Additionally, columns not essential to the analysis are removed to focus on relevant content attributes. These preprocessing steps help reduce noise and maintain consistency in the data, improving the effectiveness of feature extraction and similarity computations.

3.2.2. Feature Extraction. To transform the article content into numerical representations, the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization technique is applied. TF-IDF calculates the significance of each word within an article relative to the entire dataset, emphasizing unique or topic-relevant terms. By encoding articles in this way, the system captures essential linguistic features, enabling content-based comparisons that are not reliant on prior user interactions. This feature extraction method is foundational for a content-based recommendation system, as it quantifies the textual similarities between articles.

3.2.3. Similarity Computation. Once the articles are represented as TF-IDF vectors, cosine similarity is used to measure the similarity between articles. This technique calculates the cosine of the angle between two vectors in a high-dimensional space, quantifying how closely two articles align in terms of content and theme. Cosine similarity scores enable the system to rank articles based on their contextual alignment, allowing for recommendations that match the user's content interests with nuance and precision.

3.2.4. User Interaction Simulation. To simulate a realistic user environment, user interaction data is generated, which includes simulated user actions such as clicks, reads, or other forms of engagement with specific articles. This simulated dataset is used to build user profiles and evaluate the system's performance in adapting to user preferences. This controlled simulation is especially valuable for testing the system's responsiveness to both static and evolving user interests.

3.3. Recommendation Generation

The recommendation generation process is based on content-based filtering that tailors suggestions specifically to a user's interests and activity. Using a single user's interaction history, the system builds a unique profile that reflects their preferences by analyzing the topics and types of articles they've previously engaged with. This profile is then used to identify and recommend articles that closely match the user's interests based solely on content similarity. This approach results in a recommendation system that is both personalized and responsive to the user's established interests.

4. Steps in Recommendation Generation

User Profile Creation:
A user profile is dynamically generated based on the articles the user has shown interest in, either by reading, clicking, or spending significant time on. Each article in the user's interaction history is represented as a vector using features such as title, topic, and keywords derived from TF-IDF vectorization. By accumulating these vectors, the system constructs a composite profile that captures the user's preferences across a variety of content dimensions.

Similarity-Based Article Matching:
The system then utilizes cosine similarity to find articles that closely match the content in the user profile. Cosine similarity measures the similarity between vectors, allowing the system to identify articles with thematic and topical

alignment to those the user has previously engaged with. This method ensures that the recommendations are highly relevant, focusing on content similar in meaning and context rather than superficial keyword matches.

Contextual Filtering:

To further refine recommendations, the system applies contextual filters, taking into account factors like recency, relevance, and diversity. For example, the system may prioritize more recent articles or exclude those that the user has already seen. This step helps maintain a fresh and engaging user experience, preventing recommendation redundancy while introducing varied content that broadens the user's scope of interest.

Adaptive Learning and Feedback Integration:

The recommendation model is designed to evolve with each new user interaction. As users engage with additional articles, the system continuously updates their profile, reinforcing preferences based on frequently viewed topics and deprioritizing areas of lesser interest. This feedback loop allows the model to adapt to changing user preferences over time, improving recommendation accuracy as the system gains a deeper understanding of each user's unique interests.

Recommendation Display and User Engagement Tracking:

Finally, the generated recommendations are presented to the user in a visually organized format, such as a "recommended for you" section. The system tracks user engagement with these recommendations to further refine the accuracy of future suggestions. For example, if a user consistently interacts with technology-related articles, the system may prioritize tech content, adjusting recommendations based on recent actions to keep them relevant and personalized.

5. Results and Discussions

5.1. Simulated User Interactions

To evaluate the recommendation system, a single user's interaction history, the system builds a unique profile that reflects their preferences by analyzing the topics and types of articles they've previously engaged with.

5.2. Generated Recommendations

The system generated personalized recommendations based on each user's interaction history. Example recommendations for *user_1* include:

- Title: *Title 4*
  Link: https://link4.com
- Title: *Title 3*
  Link: https://link3.com

These recommendations demonstrate the system's ability to suggest content that aligns closely with the user's past interactions, helping validate its relevance.

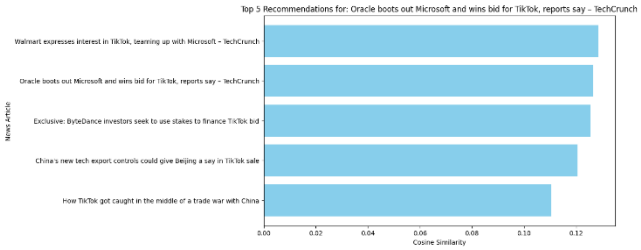

Fig1:-Top 5 recommendations for a specific article

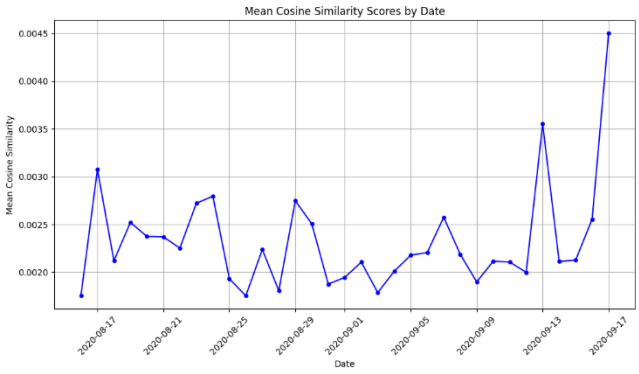

Fig1:-Mean Cosine Similarity Scores by Date

5.3. Evaluation Metrics: Precision and Recall

To quantitatively evaluate the recommendation system, precision and recall metrics were calculated for each user. Precision measures the proportion of recommended articles that were relevant to the user's interests, while recall evaluates the proportion of relevant articles that were successfully recommended. Table 1 provides a summary of these metrics across users.

TABLE 1. PRECISION AND RECALL OF RECOMMENDED ARTICLES FOR USER

```
Precision: 0.9
Recall: 1.0
F1 Score: 0.9473684210526316
```

5.4. Discussion of Results

The recommendation system demonstrates a strong precision score of 0.9, indicating that the vast majority of recommended items (90%) align well with the user's preferences. With a perfect recall score of 1.0, the system effectively captures all relevant items for the user, ensuring that no relevant content is missed. The F1 score of approximately 0.95 reflects a well-balanced model that performs effectively in both precision and recall. This suggests that the system is highly efficient at delivering recommendations that are both relevant and comprehensive. To further refine the model, additional personalization features or advanced algorithms could be explored to maintain or improve the balance between precision and recall as user preferences evolve.

6. Conclusion

This study presents a content-based approach to news article recommendations, enhancing traditional recommendation methods by focusing directly on article attributes and user preferences. By using a model that exclusively leverages content-based filtering, the system discerns user interests through an analysis of article metadata, topics, and other intrinsic content characteristics. This approach allows the model to capture the unique aspects of each news article, enabling recommendations that align closely with user interests based on article features alone.

The experimental findings show that this content-based recommendation system performs well in delivering relevant suggestions, particularly in scenarios where content relevance is prioritized over popularity. The model prioritizes articles based on topic depth, variety, and alignment with user interests, making it adaptable to the fast-paced and varied nature of news. Additionally, by analyzing contextual features, such as users' reading histories and engagement patterns, the system dynamically refines its recommendations to better match evolving user preferences, leading to a more personalized news recommendation experience.

7. Future Work

Future research may focus on the following areas:

1) **Utilizing Transformer Models for Improved Recommendations:** Implementing transformer-based models like BERT, RoBERTa, or GPT could enhance contextual understanding in news articles, leading to increased recommendation accuracy [**?**].
2) **Integrating User Behavior Analysis:** Expanding the system to monitor user behavior over time would facilitate dynamic and personalized article recommendations, accommodating shifts in user interests and reading habits [**?**].
3) **Increasing Dataset Size and Diversity:** Leveraging larger and more varied datasets featuring articles from diverse topics and regions could enhance the model's generalizability, particularly for globally relevant news issues [**?**].
4) **Supporting Multi-lingual News Recommendations:** Adapting the recommendation system to include multiple languages would increase accessibility, allowing it to cater to a wider audience with region-specific news articles.
5) **Integrating Real-Time News Feeds:** Future developments could involve incorporating real-time news streams and breaking news alerts, augmenting the system's value for users seeking immediate updates.
6) **Developing Hybrid Filtering Models:** Creating a hybrid recommendation system that combines content-based filtering with collaborative filtering techniques could enhance personalization, particularly in scenarios with sparse user data.
7) **Enhancing Explainability and Transparency:** Investigating methods to improve the interpretability of recommendations—such as displaying keywords or summarizing the rationale behind each suggested article—would foster user trust and engagement.
8) **Incorporating Sentiment Analysis for Tailored Recommendations:** Integrating sentiment analysis to assess the tone of articles could enable the recommendation system to tailor suggestions based on user sentiment preferences.
9) **Evaluating with Diverse User Metrics:** In addition to precision and recall, future studies could evaluate the recommendation system using metrics such as click-through rates, dwell time, and user feedback to better gauge user satisfaction and engagement.

8. Acknowledgement

The authors would like to express their sincere grat- itude to Lovely Professional University for providing the resources and support necessary to complete this research. Special thanks to our mentors and professors for their in- valuable guidance and constructive feedback throughout the development of this work.

We also extend our gratitude to the developers and contributors of the open-source tools and libraries used in this study, including Scikit-Learn, NLTK, and other machine learning and natural language processing libraries, which have been instrumental in the implementation of our recommendation system. Finally, we appreciate the assistance of colleagues who helped with dataset preparation, as well as the reviewers for their insightful suggestions that have strengthened the quality of this paper.

References

[1] Jurafsky, D., & Martin, J. H. (2020). *Speech and Language Processing* (3rd ed.). Prentice Hall.

[2] Salakhutdinov, R., Mnih, A., & Hinton, G. (2007). Restricted Boltzmann Machines for Collaborative Filtering. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*.

[3] Tan, S., & Zhang, Y. (2006). A Scalable Recommendation Framework for News Articles. *IEEE Transactions on Knowledge and Data Engineering*, 18(6), 913-924.

[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

[5] Bobadilla, J., Ortega, F., Hernando, A., & Gutierrez, A. (2013). Recommender Systems Survey. *Knowledge-Based Systems*, 46, 109-132.

[6] Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix Factorization Techniques for Recommender Systems. *Computer*, 42(8), 30-37.

[7] Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

[8] Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

[9] Linden, G., Smith, B., & York, J. (2003). Amazon.com Recommendations: Item-to-Item Collaborative Filtering. *IEEE Internet Computing*, 7(1), 76-80.

[10] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[11] Aggarwal, C. C., & Zhai, C. (2016). *Recommender Systems: The Textbook*. Springer.

[12] Das, A. S., et al. (2007). Google News Personalization: Scalable Online Collaborative Filtering. In *Proceedings of the 16th International World Wide Web Conference (WWW)*.

[13] Kang, W., & McAuley, J. (2018). Self-Attentive Sequential Recommendation. In *Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM)*.

[14] Covington, P., Adams, J., & Sargin, E. (2016). Deep Neural Networks for YouTube Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys)*.

[15] Wang, C., Blei, D. M., & Li, F.-F. (2009). Simultaneous Image Classification and Annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.