A Novel E-commerce Recommendation System Based on RAG and Pretrained Large Model

Guohua Xiao
School of Information Science and Technology
Sanda University
Shanghai, China
guohua.xiao@gmail.com

Jiongqian Wu
School of Information Science and Technology
Sanda University
Shanghai, China
1095936847@qq.com

Shih-Pang Tseng
School of Software and Big Data
Changzhou College of Information Technology
Changzhou, China
tsengshihpang@czcit.edu.cn

2024 International Conference on Orange Technology (ICOT) | 979-8-3315-2360-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICOT64290.2024.10936978

***Abstract*—The exponential growth of e-commerce platforms necessitates efficient and scalable recommendation systems to enhance user experience and business performance. This paper introduces a novel recommendation system based on the Retrieval-Augmented Generation (RAG) framework, leveraging customer review data as an external, modular knowledge source. Unlike traditional systems that require global computations over user-product matrices, the proposed system employs FAISS for efficient vector-based retrieval and BART for natural language generation, reducing computational overhead and enhancing scalability. The system is implemented as a web-based application offering two functionalities: checking if a specific product is recommended based on reviews and generating product recommendations from user queries. Evaluation demonstrates the system's ability to provide accurate, user-friendly recommendations while maintaining extensibility and adaptability to dynamic datasets. This work establishes a foundation for modular, review-driven recommendation systems in real-world e-commerce scenarios.**

***Keywords*—Recommendation System, Retrieval-Augmented Generation, FAISS, BART, E-commerce, Review-Based Recommendations**

I. Introduction

In recent years, the proliferation of e-commerce platforms has significantly increased the need for effective recommendation systems. These systems assist users in navigating vast catalogs of products and help businesses improve customer satisfaction and sales. Traditional recommendation systems often rely on collaborative filtering or content-based approaches, requiring global computations over user-product matrices or extensive feature engineering. While these methods are effective, they pose challenges in terms of scalability and adaptability, particularly in dynamic environments where new products and reviews are frequently added.

In light of these challenges, there is an increasing need for innovative solutions that can enhance the scalability, adaptability, and personalization of recommendation systems. The integration of novel machine learning approaches, including transfer learning, hybrid models, and reinforcement learning, presents new opportunities to improve the performance of recommendation engines, especially in the context of real-time, large-scale e-commerce platforms. Furthermore, the use of multimodal data sources, including text, images, and user behavior, offers a promising avenue for enhancing the accuracy and diversity of recommendations. This paper proposes a novel approach that leverages advanced embedding techniques and efficient indexing methods to build a scalable and adaptable recommendation system, capable of processing dynamic data streams and integrating diverse data sources for personalized recommendations in the e-commerce domain.

In detail,this paper presents a novel recommendation system based on the Retrieval-Augmented Generation (RAG) [1] framework. The proposed system leverages customer review data as an external, modular knowledge source, bypassing the need for exhaustive global computations. By integrating FAISS (Facebook AI Similarity Search) [2] for efficient vector-based retrieval and BART (Bidirectional and Auto-Regressive Transformers) [3] for natural language generation, the system offers scalable and intuitive recommendations. The approach is particularly advantageous for scenarios where review data can provide detailed insights into user preferences and product attributes.

The system was implemented as a web-based application, providing two core functionalities: (1) checking whether a specific product is recommended based on its reviews and (2) generating recommendations for products based on user queries. The use of review data as an external resource not only reduces computational overhead but also enhances the system's extensibility and modularity, allowing it to adapt to new datasets and evolving user preferences.

This paper is structured as follows: Section II reviews related work, highlighting existing approaches to review-based recommendation systems and their limitations. Section III details the methods employed, including dataset preprocessing, the architecture of the RAG framework, and the implementation of the web interface. Section IV presents the results and discussion, showcasing the system's capabilities and comparing it with traditional methods. Finally, Section V concludes with a summary of the contributions and directions for future research.

II. Related Work

Several studies have explored the use of review data in recommendation systems, focusing on the integration of textual reviews with traditional recommendation algorithms. However,

most of these approaches involve global computations on the entire user-product matrix, leading to high computational overhead. Below, we highlight three representative works and compare them with the proposed RAG-based recommendation system:

- **Review-Based Collaborative Filtering:** McAuley et al. [4] proposed a method that incorporates textual reviews into collaborative filtering models. By extracting latent factors from reviews, the system improves recommendation accuracy. However, it requires joint optimization across all users and products, resulting in significant computational demands.
- **Neural Attention Models for Reviews:** Chen et al. [5] introduced an attention-based mechanism to derive user and item representations from reviews. While effective in capturing context-specific preferences, this approach depends on training deep neural networks over the entire dataset, making it less scalable for large systems.
- **Aspect-Based Sentiment Analysis for Recommendations:** Zhang et al. [6] utilized aspect-based sentiment analysis to enhance product recommendations. This method relies on extracting and summarizing specific aspects from all reviews, which entails extensive preprocessing and aggregation efforts.

In contrast, the proposed RAG-based system leverages reviews as an external knowledge source without requiring global relevance computations. By employing FAISS for efficient vector-based retrieval and BART for natural language generation, the system achieves high scalability and modularity. This design significantly reduces computational overhead while maintaining robust recommendation performance.

III. Methods

A. Dataset

The dataset used in this study originates from a women's e-commerce clothing reviews dataset available on Kaggle at https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews. It contains detailed customer feedback on various clothing items and consists of the following columns:

- **Clothing ID:** Integer categorical variable referring to the specific piece being reviewed.
- **Age:** Positive integer representing the age of the reviewer.
- **Title:** String variable representing the title of the review.
- **Review Text:** String variable containing the body of the review.
- **Rating:** Positive ordinal integer from 1 (Worst) to 5 (Best).
- **Recommended IND:** Binary variable indicating whether the customer recommends the product (1 for recommended, 0 otherwise).
- **Positive Feedback Count:** Positive integer documenting the number of customers who found the review helpful.
- **Division Name:** Categorical variable for the high-level division of the product.
- **Department Name:** Categorical variable for the department name of the product.
- **Class Name:** Categorical variable for the product class name.

The dataset was preprocessed to fill missing values in the extitReview Text column and normalize textual data for subsequent vectorization.

B. System Architecture

The proposed recommendation system leverages the Retrieval-Augmented Generation (RAG) framework, combining retrieval mechanisms with natural language generation capabilities. The following key components were implemented(Fig.1):

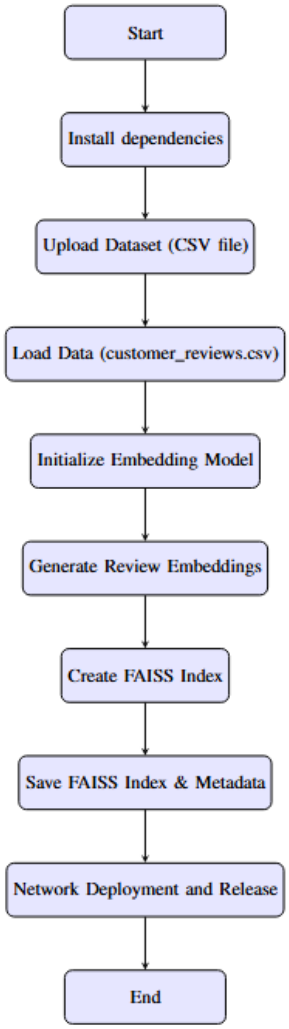

Fig. 1. The Flowchart of The Project

1) FAISS for Vector-Based Retrieval: FAISS (Facebook AI Similarity Search) is used as the vector similarity search engine. Each review is converted into a high-dimensional embedding vector using the SentenceTransformer extitall-MiniLM-L6-v2 model. These embeddings are stored in a

FAISS index, enabling efficient similarity-based retrieval for queries.

- **Embedding Generation:** The extitReview Text column is encoded into 384-dimensional vectors using Sentence-Transformer.
- **FAISS Index:** A flat L2 index is created to store these vectors, allowing for fast nearest-neighbor searches based on Euclidean distance.
- **Query Matching:** User queries are converted into embedding vectors and compared with stored embeddings to retrieve the most relevant reviews.

2) BART for Natural Language Generation: BART (Bidirectional and Auto-Regressive Transformers) is employed for generating natural language explanations based on retrieved reviews. The extitfacebook/bart-large-cnn model was fine-tuned to summarize concatenated reviews into concise, user-friendly recommendations.

- **Input:** Retrieved review texts are concatenated and prepended with a task-specific prompt.
- **Output:** The generated output provides a summarized recommendation reason or highlights key aspects of the retrieved products.

3) Web Interface: A web-based user interface was developed using Flask to facilitate user interaction. The interface integrates two functionalities:

- **Check Product Recommendation:** Allows users to input a product ID and retrieve its recommendation status and reason.
- **Query-Based Recommendations:** Enables users to input a natural language query to receive a list of recommended products and reasons.

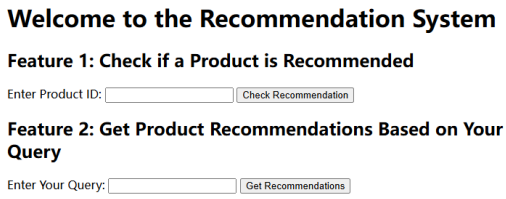

Fig. 2. The Web Interface of the Recommendation System

The interface(Fig.2) was designed for simplicity, with separate forms for each functionality. Results are displayed dynamically on new pages, enhancing usability.

IV. Results and Discussion

A. Web Interface Design

The proposed recommendation system was implemented as a web-based application to provide an intuitive and user-friendly interface for users to interact with the system. The homepage integrates two main functionalities:

- **Function 1: Check if a Product is Recommended** - Users can enter the product ID in a dedicated text box labeled *"Enter Product ID"* and submit the query by clicking the *"Check Recommendation"* button.
- **Function 2: Query-Based Product Recommendations** - Users can input a query describing their needs in a text box labeled *"Enter Your Query"* and submit the request using the *"Get Recommendations"* button.

Each functionality redirects the user to a results page, displaying either the recommendation for a specific product (Function 1) or a list of recommended products along with their reasons (Function 2). A navigation link to return to the homepage is provided for ease of use.

B. System Testing

To validate the system's functionalities, the following test cases were designed and executed using the developed web interface:

1) Function 1: Check if a Product is Recommended: A test was conducted by entering a specific product ID **853** in the *"Enter Product ID"* field and clicking the *"Check Recommendation"* button. The system successfully retrieved the relevant reviews and computed the recommendation status based on the average *Recommended IND* value. The result displayed (Fig.3):

- **Product ID:** 853
- **Is Recommended:** Yes
- **Recommendation Reason:** *"Why recommend this product? Took a chance on this blouse and so glad i did. crisp and clean is how i would describe it. fits great. drape is perfect. wear tucked in or out - can't go wrong."*

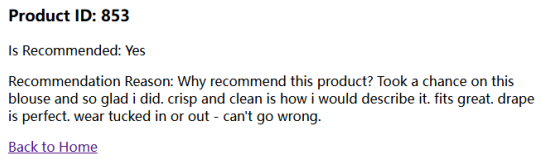

Fig. 3. Check if a Product is Recommended

This result confirms the system's ability to correctly process product-specific queries and provide a clear, concise recommendation and explanation.

2) Function 2: Query-Based Product Recommendations: The system's query-based recommendation functionality was tested by entering **"comfortable shirt"** in the *"Enter Your Query"* field and clicking the *"Get Recommendations"* button. The system performed a semantic search using the FAISS index and retrieved the most relevant products based on the query. The result displayed the following recommendations (Fig.4):

- **Product ID:** 1002 - *"Reason: Love this skirt. very comfy. looks like it's very expensive material but feels comfortable."*
- **Product ID:** 1020 - *"Reason: This skirt is really cute and very comfortable. i can dress it up for work or dress it down for casual. fun color too."*

- **Product ID:** 1001 - *"Reason: I really like the skirt! it's very comfortable and i can wear it with so many colors."*
- **Product ID:** 995 - *"Reason: Love tracy reese finds at retailer! first day wearing this skirt and have already received a number of compliments! not a typical material i would buy clothing wise, but it's really comfortable and not clingy as i thought it may be. i appreciate the built in slip for some extra substance as well"*
- **Product ID:** 1001 - *"Reason: Cute casual skirt, i love that it is longer in the back than the front so i can comfortably wear it to work. i had to size up, which i usually do in bottoms that do not have stretch because i'm my athletic frame."*
- **Summary** *"This skirt is really cute and very comfortable. i can dress it up for work or dress it down for casual. fun color too. Love tracy reese finds at retailer!"*

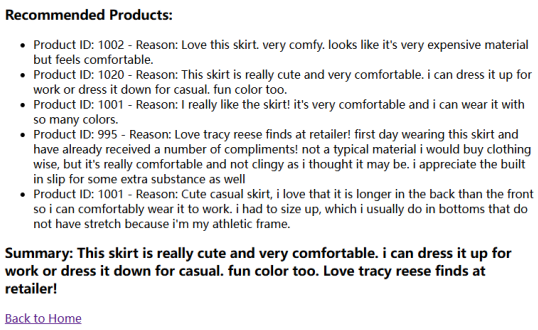

Fig. 4. Query-Based Product Recommendations

The system demonstrated its capability to match user queries to relevant products and provide detailed, review-based explanations for each recommendation. This highlights the robustness of the semantic search and recommendation logic implemented in the system.

C. Discussion

The testing results showcase the effectiveness of the proposed recommendation system in delivering accurate and user-friendly recommendations. The integration of semantic search using FAISS with the retrieval-augmented generation (RAG) framework ensures high relevance in matching user inputs to the underlying dataset. The web-based interface simplifies user interaction, making the system accessible to non-technical users.

The primary strength of the system lies in its ability to leverage review data for both product-specific and query-based recommendations. By presenting clear reasons for recommendations, the system enhances user trust and decision-making.

However, certain limitations were observed during testing. For instance, the recommendation quality depends heavily on the completeness and quality of the underlying review dataset. Products with sparse or poorly written reviews may result in less informative recommendations. Future work may focus on integrating additional data sources, such as product specifications or user preferences, to enhance recommendation accuracy.

Overall, the developed system provides a robust and scalable solution for personalized product recommendations, suitable for deployment in real-world e-commerce applications.

V. Conclusion

This paper proposes a recommendation system based on the Retrieval-Augmented Generation (RAG) framework. Unlike traditional recommendation systems that require global relevance computations for all user and product data, the proposed system leverages review data as an external, modular knowledge source. This approach significantly reduces computational overhead and enhances system scalability.

The primary contributions of this work are as follows:

- The design and implementation of a web-based RAG recommendation system that provides two functionalities: product-specific recommendation and query-based recommendations.
- Integration of FAISS for fast, similarity-based search and BART for natural language generation, enabling intuitive and meaningful recommendations.
- The use of review data as an external knowledge source, eliminating the need for exhaustive global relevance computations.

Acknowledgment

The authors thank Prof. Jhing-Fa Wang for valuable suggestions.And the project is supported by Sanda University (research project NO. 2021BSZX07).

References

[1] Lewis, Patrick, et al, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Advances in Neural Information Processing Systems*, 2020, 33.pp. 9459-9474.

[2] Krisnawati, L.D., Mahastama, A.W., Haw, S.C., Ng, K.W. and Naveen, P., "Indonesian-English Textual Similarity Detection Using Universal Sentence Encoder (USE) and Facebook AI Similarity Search (FAISS)," in *CommIT (Communication and Information Technology) Journal)*, 2024, 18(2), pp.183-195.

[3] Adhik, Chintalwar, Sonti Sri Lakshmi, and C. Muralidharan, "Text summarization using BART," in *AIP Conference Proceedings*, 2024, Vol. 3075. No. 1. AIP Publishing

[4] J. McAuley and J. Leskovec, "Hidden factors and hidden topics: Understanding rating dimensions with review text," in *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys)*, 2013, pp. 165–172.

[5] T. Chen, H. Zhang, and D. Su, "Neural attention models for recommendation," in *Proceedings of the 2018 World Wide Web Conference (WWW)*, 2018, pp. 1855–1864.

[6] Y. Zhang, Q. Yang, and J. Gao, "Aspect-based sentiment analysis for product recommendation," in *Proceedings of the 13th International AAAI Conference on Web and Social Media (ICWSM)*, 2019, pp. 600–609.