Hybrid Recommendation System

Tim Zhang Junyun Huang Jiarui Li Suzie Wei

Abstract

Recommendation systems are critical for helping users find relevant information from massive contents on digital platforms. Our project proposes a hybrid recommendation system, which combines collaborative and content-based filtering[2] , for more robust analysis and potentially increase accuracy in recommendations.

1 Motivation

With the frequent use of social media today, people now have access to millions of online data by just a few clicks. However, this also presents a new challenge: users can easily become overwhelmed by this huge amount of contents. Therefore, how to find the most relevant and useful information from the stream of data has become a commonly studied topic.

Our final project implements a hybrid news recommendation system by combining two well-known approaches: content-based filtering and collaborative filtering. This approach mitigates the limitations of each approach.

1.1 Limitations of Traditional Recommendation Systems

We will consider two traditional recommendation systems: content-based filtering (CBF) and collaborative filtering (CF). Content-based filtering works by analyzing users' interactions to match item features and generate recommendations based on similarities of itmes in user behaviors and items from the recommendation dataset. However, since CBF relies solely on users and item features relevances, this limits its ability to diversify recommendations. In other words, CBF lacks a methodology to explore the new and unpredicted (IBM). On the other hand, collaborative filtering works by categorizing users into distinct groups with common interests, then generate recommendations based on other users' behaviors in an interest group. The limitations of CF lie mainly in dealing with unique and unconventional user preferences, which makes it difficult to fit the user into any groups.

1.2 Potentials of Hybrid Recommendation Systems

Our project of combining the two traditional approaches should potentially mitigate the disadvantages of each approach and as a result improve the effectiveness of the recommendation system. For example, while CF fails to make accurate recommendations for users with very unique preferences, CBF can still recommend relevant content. In contrast, while CBF fails to diversify recommendations, using CF can help explore new recommendations by referencing users with similar preferences.

2 Proposed Approach

We are planning to train our two models with unsupervised learning on Microsoft News dataset MIND[1] to develop a hybrid recommendation system aimed at delivering personalized article recommendations by combining content-based filtering (CBF) and collaborative filtering (CF). Our approach leverages two models to balance exploration and exploitation that enhances both diversity and accuracy.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

2.1 Content-based Filtering model

The first model will analyze users' view histories, generating recommendations by matching with similar contents. This approach ensures that users receive recommendations that align closely with their demonstrated interests, thereby supporting the exploitation aspect by focusing on personalized content.

2.2 Collaborative Filtering Model

The second model will identify similar user profiles and recommend popular articles within these peer groups. By analyzing the preferences of users with comparable interests, this model introduces an exploration element, providing users with new content they may not have discovered independently but users in their belonging peer groups are interested in.

Each model will output a ranked list of articles scored based on relevance, and the system will dynamically balance the contributions of both models to optimize for a blend of familiarity and discovery in the recommendations.

3 Expected plan and moonshot goals

Our recommendation system will consist of mainly 4 components: data processing module, CBF model, CF model, and integrated recommendation module. We plan to first finish data processing in 1 2 weeks, then 1 week for each model, and lastly the recommendation in <= 1 week.

3.1 Data

Data Preparation and Preprocessing on MIND: we will start with data cleaning, preprocessing, and feature extraction to ensure compatibility with both CBF and CF models. This includes parsing user interaction data, standardizing article metadata, and engineering features suitable for the system.

3.2 Training CBF Model

We will develop a CBF model that recommends articles based on similarity in article features, such as topic, keywords, and other metadata.

3.3 Training CF model

In parallel, a CF model will be developed to analyze user profiles and recommend articles based on peer group preferences.

3.4 Integration

Once the two models are trained, we will integrate them into a hybrid system, dynamically adjusting the weighting of CBF and CF recommendations to achieve a balance between familiarity (exploitation) and discovery (exploration). This balancing mechanism will be tuned to optimize user engagement.

3.5 Evaluation and Metrics

Key metrics for evaluation include:

Accuracy: Ensuring the relevance of recommendations by measuring metrics like precision and recall. Diversity: Using diversity scores to assess how well the system introduces users to new, unexplored content. User Satisfaction: Survey-based feedback or implicit user engagement metrics (e.g., click-through rate) to gauge user satisfaction and acceptance.

3.5.1 Moonshot Goals

To push the boundaries of conventional recommendation systems, we aim for several ambitious goals:

Real-Time Personalization: Incorporate real-time feedback loops that adjust recommendations based on users' immediate interactions. This will require optimization for low-latency responses and adaptive learning.

Cross-Platform Recommendations: Expand the recommendation system's applicability by tailoring the model to different platforms beyond news, such as social media, e-commerce, or educational content. This would allow the system to serve a broad range of industries and enhance its versatility.

Explainability and Transparency: Develop a transparent interface that explains the recommendation process to users, enhancing trust and providing insights into why specific content is suggested. This would involve building interpretable models or layers that highlight key features influencing recommendations. . Through these goals, we aspire not only to refine hybrid recommendation systems but also to pioneer innovative approaches that enhance user experiences across digital platforms.

4 Methods

We trained an embedding model to generate embeddings for news information extracted from Microsoft's MIND News dataset. And then use the embeddings to compare for relevance between different news articles. And apply these similarities in CBF and CF respectively for recommendations.

4.1 Dataset Overview and Processing

The MIND dataset was used for training our embedding model, which includes a news.tsv file containing structured news data with the following fields: id, category, subcategory, title, abstract, and entities. For model training, we preprocessed it by combining the category, subcategory, title, and abstract fields into a single text field, referred to as "news info." Subsequently, any entries where the combined "news info" exceeded a length of 512 characters were dropped to ensure compatibility with the input length constraints of the model.

4.2 Training the Embedding Model

We trained our embedding model on a base model, BERT-base-uncased [5], to generate contextual embeddings in the news domain on our preprocessed dataset described above. The training process followed an unsupervised Masked Language Modeling (MLM) approach [3], where random tokens in the input text were masked, and the model was tasked with predicting the original tokens.

We used two types of data collators: DataCollatorForWholeWordMask for whole-word masking and DataCollatorForLanguageModeling for standard token-level masking, depending on whether whole-word masking was enabled. The masked tokens were selected with a probability determined by the specified mlm probability. Inputs were tokenized and inputs with sequences larger than 512 tokens are dropped considering the compactability with our embedding model.

We used Transfomer's API to help handle the training process, managing the dataset, masking logic, and model updates. Configuring training arguments to specify the number of epochs, batch size, evaluation frequency, and other key parameters. After training, the model was saved to the output directory, providing embeddings that capture semantic and contextual relationships from news data.

4.3 Content-Based Filtering for News Recommendation

To personalize news recommendations, we implemented a Content-Based Filtering (CBF) module that identifies the most relevant articles based on a user's reading history. Each user's historical interactions were represented by extracting semantic embeddings of previously read articles using a transformer-based model. To enhance computational efficiency, articles in the dataset were categorized by subcategory, limiting comparisons to relevant subsets. Cosine similarity was utilized to rank articles by relevance, and the top results were recommended. This approach ensured that the recommendations were both personalized and contextually aligned with user preferences, significantly improving the system's relevance and engagement.

4.4 Collaborative Filtering for News Recommendation

The collaborative filtering (CF) model was a critical component of the hybrid recommendation system, designed to analyze user interaction data and identify patterns of shared interests among users. To achieve this, clustering techniques were employed to group users based on similar interaction patterns, enabling the model to effectively capture collaborative aspects of recommendations. Recommendations were generated by identifying popular articles within each peer group, introducing users to content aligned with their interests and enhancing the exploration aspect of the system. To further improve recommendation diversity, the CF model incorporated diverse yet popular articles from peer groups, balancing familiarity (exploitation) and discovery (exploration). This approach allowed the system to deliver personalized recommendations tailored to user preferences while encouraging the exploration of new and diverse content. By combining these techniques, the CF model significantly enhanced the system's ability to adapt to unique and unconventional user behaviors, resulting in improved overall performance.

5 Experiments

5.1 Training Phase

We experimented with different base models, for example BERT base-uncased, BERT-base-large, Roberta base, etc. And hyperparamter tuned on them to get a relatively optimal loss for both train and validation sets. We also compared the similarity scores between embeddings, which showed that training with our news data helped finding more relevant and less relevant contents. As more relevant data got higher socres and less relevant got lower socres.

Figure 1 shows Roberta-base and Figure 2 shows BERT-base [4] train and val loss.

Step	Training Loss	Validation Loss
1000	2.214600	1.839947
2000	1.949800	1.740839
3000	1.851000	1.669927
4000	1.765800	1.619567
5000	1.746900	1.593121
6000	1.701000	1.563355
7000	1.677000	1.544158
8000	1.634400	1.530109
9000	1.617100	1.507897

Figure 1: Roberta-base train and validation loss

Step	Training Loss	Validation Loss
1000	2.221500	1.876811
2000	2.003200	1.775177
3000	1.898000	1.693791
4000	1.785800	1.637199
5000	1.753900	1.590190
6000	1.697300	1.546615
7000	1.648300	1.516157
8000	1.588900	1.488654
9000	1.562000	1.458267

Figure 2: bert-base-uncased train and validation loss

5.2 Collaborative Filtering Testing

The Collaborative Filtering test results revealed that users within closely related groups exhibited similar reading preferences, frequently engaging with the same news articles. The CF model effectively leveraged these patterns, recommending commonly shared articles to other users within

```
News ID: N55189, Category: tv, Title: 'Wheel Of Fortune' Guest Delivers Hilarious, Off The Rails Introduction
News ID: N42782, Category: sports, Title: Three takeaways from Yankees' ALCS Game 5 victory over the Astros
News ID: N34694, Category: tv, Title: Rosie O'Donnell: Barbara Walters Isn't 'Up to Speaking to People' Right Now
News ID: N45794, Category: news, Title: Four flight attendants were arrested in Miami's airport after bringing in thousands in cash, police say
News ID: N18445, Category: sports, Title: Michigan sends breakup tweet to Notre Dame as series goes on hold
News ID: N63302, Category: lifestyle, Title: This Wedding Photo of a Canine Best Man Captures Just How Deep a Dog's Love Truly Is
News ID: N10414, Category: movies, Title: Robert Evans, 'Chinatown' Producer and Paramount Chief, Dies at 89
News ID: N19347, Category: news, Title: Former US Senator Kay Hagan dead at 66
News ID: N31801, Category: news, Title: Joe Biden reportedly denied Communion at a South Carolina church because of his stance on abortion
```

Figure 3: Sample of user's view history.

```
News ID: N18445, Category: sports, Title: Michigan sends breakup tweet to Notre Dame as series goes on hold
News ID: N34694, Category: tv, Title: Rosie O'Donnell: Barbara Walters Isn't 'Up to Speaking to People' Right Now
News ID: N42782, Category: sports, Title: Three takeaways from Yankees' ALCS Game 5 victory over the Astros
News ID: N45794, Category: news, Title: Four flight attendants were arrested in Miami's airport after bringing in thousands in cash, police say
News ID: N55189, Category: tv, Title: 'Wheel Of Fortune' Guest Delivers Hilarious, Off The Rails Introduction
```

Figure 4: Recommendation results from Collaborative Filtering.

the same group. This demonstrated the model's ability to identify and utilize group-based interests to enhance the relevance and personalization of its recommendations.

5.3 Content-Based Filtering Testing

```
11
Article ID: N30924
Title: The Rock's Gnarly Palm Is a Testament to Life Without Lifting Gloves
Top 5 Most Relevant Articles:
ID: N44111, Title: 14 Celebs Over 50 Who Are In The Best Shape Of Their Lives, Similarity: 0.8807
ID: N9140, Title: 85-Year-Old Runner Who Keeps Racking Up Medals: 'It's My Job to Get People Moving', Similarity: 0.8735
ID: N58885, Title: How This Guy Built Muscle and Got Shredded in His 40s on a Plant-Based Diet, Similarity: 0.8735
ID: N53700, Title: Ellen DeGeneres Likes to Do This Quick Abs Workout Before Hitting the Stage for Her Talk Show, Similarity: 0.8707
ID: N44005, Title: How David Boreanaz Stays 'Seal Team' Fit at Age 50, Similarity: 0.8691
------------------------------------------------------------
100%|██████████| 1/1 [00:01<00:00,  1.33s/it]12
Article ID: N6778
Title: A part of Windsor Castle that's been closed to the public for over 150 years just reopened   and it's been completely revamped. Take a peek inside.
Top 5 Most Relevant Articles:
ID: N8071, Title: The Brands Queen Elizabeth, Prince Charles, and Prince Philip Swear By, Similarity: 0.8456
ID: N19818, Title: Kate Middleton's Best Hairstyles Through the Years, Similarity: 0.8450
ID: N1810, Title: Every outfit Duchess Kate has worn in 2019, Similarity: 0.8392
ID: N53921, Title: 25 Photos of the Royal Family at Balmoral Castle, Queen Elizabeth's Favorite Home, Similarity: 0.8344
ID: N63495, Title: Prince George's Royal Life in Photos, Similarity: 0.8223
------------------------------------------------------------
```

Figure 5: Experimental results of the Content-Based Filtering module. For each article in a user's reading history, the system retrieved the top 5 most relevant articles with similarity scores.

Figure 5 illustrates the output of the Content-Based Filtering module for two articles from a user's reading history. For each base article, the system identified the top 5 most relevant articles from the dataset based on cosine similarity scores.

The results demonstrate the CBF module's ability to effectively identify and recommend articles that align closely with the user's interests. For example:

- For the article *"The Rock's Gnarly Palm Is a Testament to Life Without Lifting Gloves"*, the module recommended articles related to fitness and celebrity workout routines with high similarity scores (e.g., 0.8807, 0.8735).
- For the article *"A Part of Windsor Castle That's Been Closed to the Public for Over 150 Years Just Reopened"*, recommendations included articles about royal family history and related events, with similarity scores as high as 0.8456.

This experiment highlights the system's ability to recommend contextually relevant articles, providing users with both familiar and new content while maintaining personalization.

6 Conclusion

We aimed to develop the hybrid recommendation system that leverages the strengths of both content-based filtering and collaborative filtering to provide more personalized and diverse recommendations. As shown by the examples above, our CBF did provide closly relevant recommendations, while CBF gave more diverse recommendations. By addressing the limitations of each individual approach, the hybrid system offers a more robust solution for recommendation tasks.

Future work could explore integrating contextual data, such as time and location, to further refine recommendations. Additionally, there are other unsupervised approach such as TSDAE, SimCSE,

GenQ, etc. And due to the structure of our dataset and time issue, we were not able to test training the model with supervised learning, which may result in better embedding models.

7 References

[1] Microsoft News Dataset (MIND). Available: https://msnews.github.io/

[2] IBM, "What is a recommendation engine?" Available: https://www.ibm.com/think/topics/recommendation-engine

[3] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks," *arXiv preprint arXiv:2004.10964*, 2020. Available: https://arxiv.org/abs/2004.10964

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2018. Available: https://arxiv.org/abs/1810.04805

[5] Google Research, "BERT Model GitHub Repository." Available: https://github.com/google-research/bert