COMP 9727 Recommender Systems
Project Design Report
MIND Microsoft News Dataset

Z5352294 Linbo Zhang

Submission Date: 30 June 2024

Contents

1. Project Scope

1.1 Project Background

Today, the news consumption has shifted dramatically from traditional print media to online platforms. In the past, traditional print media decided what news that we read. This news is often printed in magazines or played on the TV. We read what they feed us. However, in today, with the vast amount of news articles published daily, as well as the arising usage of mobile phones and laptops, the users start to decide what they want to read, by a simple click on the news. And the news media face significant challenges in increasing the click rate of their news. The reason is that the click rate is usually related to the advertising. These ads are embedded in the news article. More clicks mean more income. And the news media starts to face significant challenges in finding relevant and interesting content. Personalized news recommendation systems aim to solve this problem by filtering and suggesting articles that match users' preferences and reading habits.

However, we need to recognize some key differences between news recommendations with other popular recommendations topics such as the movie representation. For example, news articles on news websites update very quickly, while a famous movie on the IMDB may last for 50 or more years. News articles are posted continuously, and it has a strong time trend. For example, when it is close to the American President election, then there will be more election news. And after the election, such news will disappear and be replaced by other new trending topics. Secondly, the cold-start problems are very severe in the news recommendation. They last short, hence the system needs to prompt them harder. Thirdly, there is no explicit rating of news articles. People will not click on the 'LIKE' button on the news article as they do on Instagram. As a result, the only reliable is whether the user clicks on the article to read the body or not.

The Microsoft News Dataset (MIND) is a large-scale dataset for news recommendation research. It was released by Microsoft in 2000. It is designed to facilitate the study of personalized news recommendation systems. Microsoft held a research competition based on the dataset. And since then, there are already 133 papers regarding varying aspects of the dataset as well as varying methods of the recommendation approach.

1.2 Project Domain and Intended Users

The domain of the recommendation system is online news. If we narrow down the definition, we are talking about digital news platforms such as Microsoft News, Google News, Yahoo News, etc. The intended users are individuals who consume news regularly and prefer to have personalized content delivery based on their reading habits

and preferences. These users often read news on their mobile devices or on their laptops. The system aims to enhance user experience by recommending news articles that align with their interests, thereby increasing the user engagement, satisfaction, and loyalty.

1.3 Number of Items Presented, User Interface and User Feedback

The use interface of the recommender system mainly focuses on the front page, which is the page that users are first exposed to when they open the news app or open the website on the laptop. The design of the front page aims to present news in a visually appealing and accessible manner. On the top of the page, we will see the website banner. It will broadcast some top news stories of the day. Often the news in the banner includes a very large title and engaging images. Below the banner, we often see news classified into several categories. For each category, we will see the news title and abstraction of some key words for that news.

Such a home page will contain a large amount of news. From the dataset analysis, there is an average of 37.4 news articles per 'impression'. The term 'impression' refers to a single instance where a user is presented with a collection of news articles. This will occur when the reader visits the front page or refreshes the front page. Each impression will record the news ID of the news presented, as well as whether the user clicks on the news or not.

The main feedback is whether the user clicks on the news or not. There is no explicit rating of news articles. The common method is usually inferred from the click behaviors in an implicit way.

1.4 Business Model and Revenue Generation

The business model for such a system will focus on generating revenue through subscription, advertising, and data insights and analytics.

Platforms like The Wall Street Journal provide subscriptions to deliver high-quality news and insights. The benefit of the subscription is that users can enjoy an ad-free experience and exclusive contents. In comparison, platforms like Google News and Microsoft News are targeting the general public. They embed advertisements within their news content to generate revenue. Those ads are not randomly generated. They are personalized so that those ads often have a higher click rate. Furthermore, offering data insights and analytics services to publishers and marketers will provide another significant revenue stream. By analyzing user engagement and content performance,

the platform can deliver valuable insights that help stakeholders optimize their strategies.

In the project, the news will be free on the website, same as what happens on Microsoft News. And the revenue will be generated from the advertising and data insights and analytics.

2. Dataset

2.1 Background and Data Structure

The MIND dataset includes the behaviour logs of 1 million users who had at least 5 news clicks during a six-week period from October 12 to November 22 in 2019. Each user has been delinked from the production system and only their hash ID is included in the dataset.

- The 6th week data is used for test.
- The 5th week data is used for training.
- The data from the last day of the 5th week is used for validation.

The dataset has been split already. In each folder, it contains the following 4 files,

- behaviors.tsv with the click histories and impression logs of users.
- news.tsv with the information of news articles. The news URL is also included, and we can scrape the news body and put into the modelling. However, since the news are at least 4 years ago, some URLs already expire, which means the experiments may not be able to work on the news body anymore. This should raise little concern, as when the user opens the front page, the user mainly sees the news title, abstraction, or keywords and then decide whether to click to view more or not.
- entity-embedding.vec and relation_embedding.vec. These two files contain the 100-dimensional embeddings of the entities and relations learned from the WikiData knowledge graph. In simpler terms, the embeddings represent the vector of that news on a large knowledge graph. These two files can facilitate the research of the knowledge-aware news recommendation.

2.2 Preliminary Analysis

The number of records of the dataset is summarized in the Table 1.

Table 1 Number of samples in dataset.

Dataset	behaviors.tsv	news.tsv
Training set	2,232,748	711,222
Validation set	376,471	72,023
Test set	2,370,727	120,959

There is a total of 1 million user records and a total of 161013 news articles. From previous paragraph, the dataset records users who have at least 5 news click records in that 6-week period. Hence duplicate use ID is common in the dataset.

Figure 1 shows a screenshot of the behaviors.tsv file. In the history column, each ID represents a piece of news. And in the impressions column, each ID represents a news and a click status, which 1 means the news is clicked, and 0 means not. Each row represents the click status after the user is feed with the 'history'.

	Impression ID	User ID	Time	History	Impressions
0	1	U87243	11/10/2019 11:30:54 AM	N8668 N39081 N65259 N79529 N73408 N43615 N2937...	N78206-0 N26368-0 N7578-0 N58592-0 N19858-0 N5...
1	2	U598644	11/12/2019 1:45:29 PM	N56056 N8726 N70353 N67998 N83823 N111108 N107...	N47996-0 N82719-0 N117066-0 N8491-0 N123784-0 ...
2	3	U532401	11/13/2019 11:23:03 AM	N128643 N87446 N122948 N9375 N82348 N129412 N5...	N103852-0 N53474-0 N127836-0 N47925-1
3	4	U593596	11/12/2019 12:24:09 PM	N31043 N39592 N4104 N8223 N114581 N92747 N1207...	N38902-0 N76434-0 N71593-0 N100073-0 N108736-0...
4	5	U239687	11/14/2019 8:03:01 PM	N65250 N122359 N71723 N53796 N41663 N41484 N11...	N76209-0 N48841-0 N67937-0 N62235-0 N6307-0 N3...

Figure 1 First 5 rows of the behaviors.tsv file in the training set.

Figure 2 shows a screenshot of the news.tsv file. For each news, there is a main category, subcategory, title, abstract, and URL. Then there is a title entity and an abstract entity that consists of more keys. The entity_embedding.vec and relation_embedding.vec maps these entities into embedding vectors.

	News ID	Category	Subcategory	Title	Abstract	URL	Title Entities	Abstract Entities
0	N88753	lifestyle	lifestyleroyals	The Brands Queen Elizabeth, Prince Charles, an...	Shop the notebooks, jackets, and more that the...	https://assets.msn.com/labs/mind/AAGH0ET.html	[{"Label": "Prince Philip, Duke of Edinburgh",...	[]
1	N45436	news	newsscienceandtechnology	Walmart Slashes Prices on Last-Generation iPads	Apple's new iPad releases bring big deals on l...	https://assets.msn.com/labs/mind/AABmf2I.html	[{"Label": "IPad", "Type": "J", "WikidataId": ...	[{"Label": "IPad", "Type": "J", "WikidataId": ...
2	N23144	health	weightloss	50 Worst Habits For Belly Fat	These seemingly harmless habits are holding yo...	https://assets.msn.com/labs/mind/AAB19MK.html	[{"Label": "Adipose tissue", "Type": "C", "Wik...	[{"Label": "Adipose tissue", "Type": "C", "Wik...
3	N86255	health	medical	Dispose of unwanted prescription drugs during ...	NaN	https://assets.msn.com/labs/mind/AAISxPN.html	[{"Label": "Drug Enforcement Administration", ...	[]
4	N93187	news	newsworld	The Cost of Trump's Aid Freeze in the Trenches...	Lt. Ivan Molchanets peeked over a parapet of s...	https://assets.msn.com/labs/mind/AAJgNsz.html	[]	[{"Label": "Ukraine", "Type": "G", "WikidataId...

Figure 2 First 5 rows of the news.tsv file in the training set.

Figure 3 to Figure 5 are generated on the training set. From these plots, we can obtain some key insights about the news, and understand why they are a completely different topic compared to other recommendation fields.

Figure 3 shows the distribution of news title lengths in the training site. The plot indicates that most news titles are concise, with a peak around 9 – 10 words. This indicates that users are often exposed to brief and to-the-point headlines. That means, a striking topic will help to increase the clicking rate. The relatively short length of news

titles suggest that the model must be adopt at extracting meaningful information from concise texts.

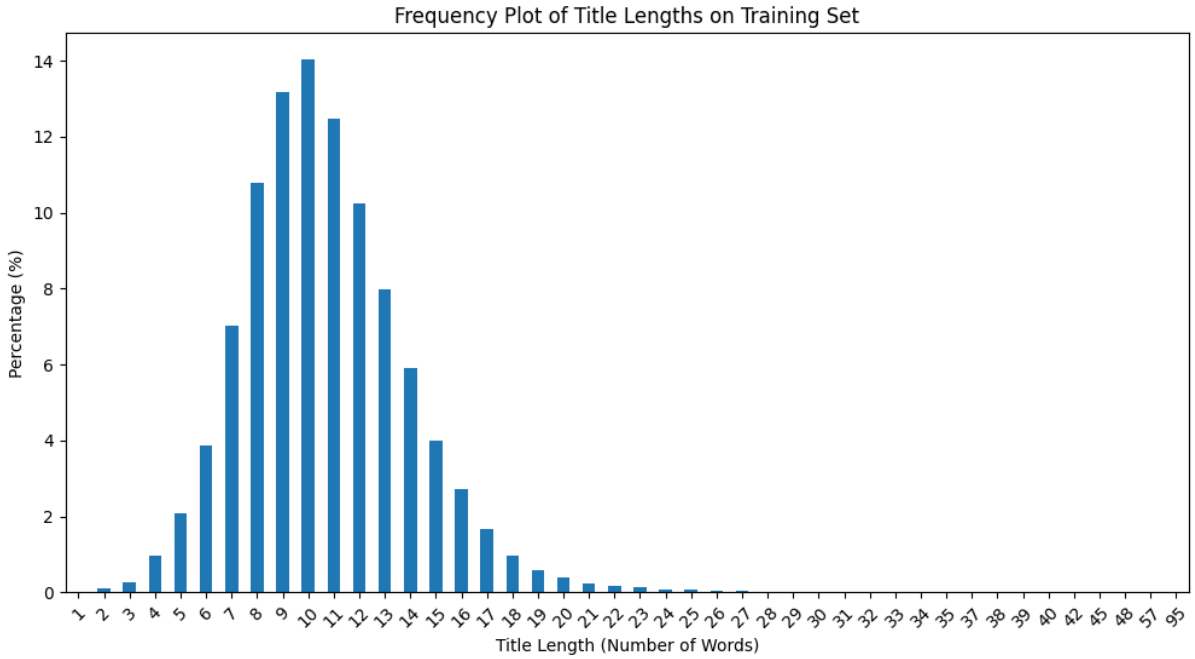

Figure 3 News title length distribution in training set.

Figure 4 illustrates the distribution of abstract lengths for new articles in the training set. There are two peaks on the graph. The first leak is around 20 words. And the second peak is around 70 words. The bimodal distribution suggests that news abstracts vary significantly in length, possibly due to the different categories or article types. The presence of both short and long abstracts means the model must handle varying amounts of text information.

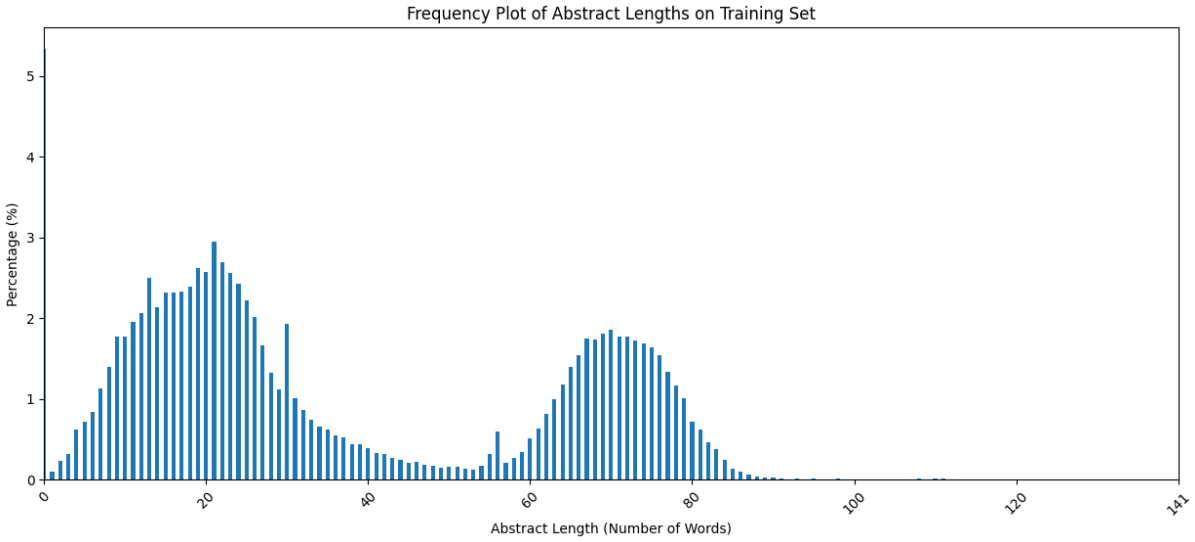

Figure 4 News abstract length distribution in training set.

Figure 5 shows the article survival time on the training set. I split the impression column and link each news ID in the column with the datetime of the impression. Then for each news ID I extract the first and last datetime and calculate the difference in days. We can consider this plot as a mimic of the survival time of each news. The plot indicates that a majority of news articles (nearly 70%) receive clisk only on the day they are published, with a sharp decline in the percentage after one day. This rapid decay is a common example of Poisson distribution. And it highlights the dynamic nature of news consumption and the cold-start problem associated with the news recommendation models. Hence, the model needs to better prompt the news before the news utility expires.

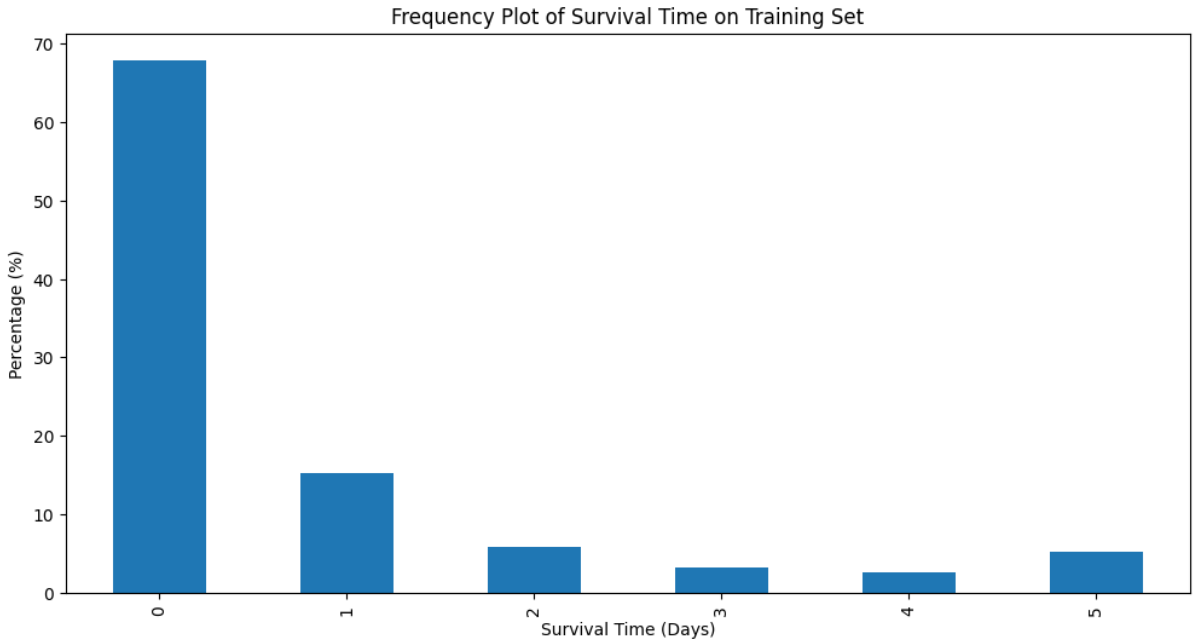

Figure 5 News survival time distribution in training set.

2.3 Challenges in News Recommendation

There are some important features of the news recommendation that requires attention during the research.

- News articles typically have a very short lifespan (cold start problem). So, the model needs to continuously update and adapt to the latest news to provide timely recommendation.
- The model needs to 'understand' the content of the news.
- Users may exhibit different reading behaviours based on the length and type of news content.

3. Methods

In the project, I plan to practice with two models,

1. A content based recommender system that uses attention mechanics and transformers
2. A hybrid recommender system that combines the content-based filtering with collaborative filtering with advanced models such as transformers.

3.1 Model 1: Content-Based Recommender System Using Attention and Transformers

This model focus on understanding and leveraging the textual content of news articles to make recommendations. The reason is that each news article contains rich textual information that can be used to infer user preferences. This model is effective in mitigating the cold start problem for new articles, as recommendations are based on content rather than historical interaction data. In addition, it will be suitable for a dynamic environment where news articles are constantly being published.

1. For Text Representation:
   Use various text representation methods to capture the semantic meaning of news articles. These methods include:
   - TF-IDF (Term Frequency-Inverse Document Frequency): A traditional method that weighs the importance of words in a document relative to their frequency in the corpus.
   - Word Embeddings (e.g., GloVe): Pre-trained word embeddings that capture semantic relationships between words.
   - Advanced NLP Models (e.g., BERT): Fine-tune pre-trained models like BERT on the news dataset to capture nuanced textual representations. This approach will be based on whether our group can access graphical cards with larger RAM.
2. News Representation Models:
   Try the following approaches:
   - Long Short-Term Memory Networks (LSTMs): Capture long-range dependencies and context within the text.
   - Multi-Head Self-Attention and Transformers (e.g., NRMS): Capture relationships between words and phrases more effectively using attention mechanisms and transformers.
3. User Profile Modelling:

Construct user profiles based on the textual content of previously clicked news articles.

- Averaging Word Embeddings: Simple yet effective way to create user profiles by averaging the embeddings of clicked articles.
- Attention Mechanisms: Prioritize more informative and relevant articles in user profiles using attention scores.
- Transformers: Use transformers to model user profiles by capturing complex dependencies between clicked articles. (This depends on whether our group can access large RAM graphical card)

4. Recommendation
   We can either match the user profile with the candidate news articles based on content similarity or use similarity measures like cosine similarity to rank and recommend articles.

3.2 Model 2: Hybrid Recommender System (Content-Based with Collaborative Filtering Using Transformers)

The hybrid recommender system is proposed to leverage the strengths of both approaches. It aims to enhance recommendation accuracy by incorporating user interaction data along with textual content. And it helps to adapt with different user preferences when we balance the user preferences and news articles content.

1. Content-Based Component
   This part will use the same text representation and user profile modelling techniques as described in the model 1. It can generate initial recommendations based on content similarity between user profiles and candidate articles.
2. Collaborative Filtering Component
   Build collaborative filtering methods to capture user-user and item-item interactions. We can consider the following methods:
   - Matrix Factorization (e.g., SVD): Decompose the user-item interaction matrix to uncover latent factors representing user preferences and item characteristics.
   - Neural Collaborative Filtering: Use neural networks to model complex interactions between users and items.
   - Transformers: Use transformers to capture complex patterns in user-item interactions.
3. Combining Methods
   We can test the following combining approaches, including
   - Weighted Hybrid: Assign weights to the scores from content-based and collaborative filtering components and combine them to generate final recommendations.

- Switching Hybrid: Switch between content-based and collaborative filtering methods based on the availability of interaction data (e.g., use content-based for new users and collaborative filtering for existing users with sufficient interaction data).

4. Evaluation

For the evaluation metrics, we will follow the common metrics used by the papers of the dataset, including AUC, MRR, nDCG@5 and nDCG@10.

- AUC (Area Under the Curve):
  AUC measures the ability of the model to distinguish between clicked and non-clicked articles. It evaluates the probability that a randomly chosen clicked article is ranked higher than a randomly chosen non-clicked article. A higher AUC indicates better performance in ranking relevant articles higher.

- MRR (Mean Reciprocal Rank):
  MRR calculates the average of the reciprocal ranks of the first relevant item in the recommended list. If the first relevant item is ranked at position k, the reciprocal rank is $1/k$. This metric is important for evaluating the accuracy of the top recommendation. It reflects how quickly the model can present a closely related article.
- nDCG@5 and nDCG@10 (Normalized Discounted Cumulative Gain):
  nDCG measures the ranking quality by considering the position of relevant articles in the recommended list. nDCG@5 and nDCG@10 focus on the top 5 and top 10 recommendations, respectively. These attributes evaluate the model's ability to place relevant articles at the top of the recommendation list. A higher result indicates a better performance in prioritizing relevant articles.

In addition, our group will also focus on the time duration for the recommendation system to provide the results, as a longer front page loading time may significantly decrease user experience.

5. Reference

Msnews (2021) *MIND: MIcrosoft News Dataset*, *MIND*. Available at: https://msnews.github.io/ (Accessed: 30 June 2024).

Msnews: *Msnews/msnews.github.io*, *GitHub*. Available at: https://github.com/msnews/msnews.github.io/tree/master (Accessed: 30 June 2024).

Papers with code - mind dataset (no date) *MIND Dataset | Papers With Code*. Available at: https://paperswithcode.com/dataset/mind (Accessed: 30 June 2024).