Group 49: Feedback-based news recommendation system

REPORT

Sakshi Sinha (MT22121) Shambhavi Pathak (MT22067) Debnath Kundu (MT22026) Snehal Buldeo (MT22074)

Problem Statement:

The project's objective is to create a new recommendation system (**NRS**) that can identify bias in suggested news items, notify readers who could be at risk of entering a ***filter bubble,*** and recommend news from a diverse topic if they want. To mitigate the filter bubble, the system should be able to identify when a user is reading a majority of news articles that are biased towards a certain topic or theme and allow them to break out of it. If the user chooses to do so, the system should recommend random articles that are more diverse in perspective and less biased. Filter Bubble refers to a situation where people are shown a narrow and personalized content selection based on their past behaviour, preferences, and interests. In the context of news recommendations, users are less likely to encounter news stories outside of their interests.

The risk of entering a filter bubble is due to a personalized recommendation system that certain applications tend to use to provide a better user experience by providing the users with news articles tailored to their individual preferences and interests. In contrast, there is a responsible news recommender system that is designed to prioritize the accuracy, credibility, and diversity of news articles, to provide users with a more balanced and trustworthy news experience. Responsible news recommenders work by attempting to surface news articles that represent a diverse range of perspectives and viewpoints unknown to a user. With our work, we want to alert the reader about the biases in their reading and recommend diverse topics with their consent. The existing applications notify the users about the bias in the news or of the news source. On the other hand, we tend to identify the biases in a user's reading pattern by collecting implicit feedback such as clicks.

Updated Literature Review:

The filter bubble phenomenon is a common concern in news recommendation systems and it occurs when the system narrows the information and deprives users of diverse information. Once the recommendations have been made, they need to be evaluated, to check the diversity in the recommended items. A diversified list more likely contains the user's actual search intent [8]. Despite the rise of interest and work on the topic in recent years, we find that a clear common methodological and conceptual ground for the evaluation of these dimensions is still to be consolidated [9]. To measure the diversity of a recommendation list, a common evaluation measure is the average pairwise distance between items in the ranked list. This measure is called intra-list diversity (ILD) and a high value in ILD means the recommended list contains items with a broad range of content [8]. introduced a rank and relevance-sensitive intra-list diversity measure (RR-ILD) that shows to what extent the recommender can diversify the
list and preserve the relevant items in the high ranks. The measures used to calculate diversity for a list of items depend on the type of recommendation system.
As we want to introduce diversity into an already recommended list of articles, it focuses on one user and one set of recommended articles. These articles are checked for diversity based on their content and headline.

Dataset Description:

News Category Dataset:

1. The dataset contains news articles from the Huffington Post, published between 2012 and 2018.
2. The dataset is available in JSON format and contains 200,853 articles.
3. Each article is represented as a dictionary in the JSON file, with the following key-value pairs:
    - **authors**: A string representing the authors of the article.
    - **category**: A string representing the category of the article.
    - **date**: A string representing the date when the article was published.
    - **headline**: A string representing the headline of the article.
    - **link**: A string representing the link to the article.
    - **short_description**: A string representing a short description of the article.
    - **text**: A string representing the content of the article.
4. The **category** field contains 41 unique categories, including politics, business, entertainment, crime, sports, etc.
5. The dataset is commonly used for text classification and natural language processing tasks, including sentiment analysis, topic modelling, and text summarization.
6. In our project, the dataset is used to train and test the news recommendation system using a content-based approach. The article headlines and content are preprocessed and transformed into feature vectors using the TF-IDF vectorization technique. The cosine similarity between articles and user profiles is used to make recommendations.

This dataset is suitable for building a basic news recommendation system. Here are some reasons why:

1. **Variety of news categories**: The dataset contains news articles from different categories such as business, entertainment, politics, sports, etc. This variety of categories ensures that the system can recommend articles from different topics to cater to the diverse interests of users.
2. **A large number of articles**: The dataset contains over 200,000 news articles, which is a large enough sample size to train the recommendation system and make accurate predictions. The more data available, the more the system can learn from the patterns in the data and make better recommendations.
3. **High-quality data**: The dataset appears to be of high quality, with well-formed and structured data, and few missing values. This quality of data is important as it ensures that the recommendation system can learn from the data effectively and make accurate predictions.
4. **Text-based content**: The dataset contains text-based content such as article headlines and article descriptions, which makes it suitable for a content-based recommendation system that relies on the text features of the articles to generate recommendations.

Methodology (Baseline) :

The pipeline for the baseline model consists of four main steps, as follows:

1. Data collection: The first step is to collect a dataset of news articles. In our project, news category dataset is used.

2. Data preprocessing: The raw text data in the dataset must be preprocessed to make it suitable for analysis. The following preprocessing steps are performed:

    - Removing punctuation and special characters
    - Converting all text to lowercase
    - Removing stop words
    - Stemming or lemmatizing the text

3. Create a document-term matrix: Represent each article in the dataset as a numerical vector using Bag-of-Words (BOW) or Term Frequency-Inverse Document Frequency (TF-IDF) methods. This will create a document-term matrix.

4. Calculate similarities between articles: Use pairwise distance similarity measure to calculate the similarity between each article in the dataset.

5. Recommend similar articles: Given a new article as input, calculate its similarity with each article in the dataset and recommend the most similar articles. Here, both BOW and TF-IDF methods have been used for news recommendations.

Future Work:

1. **Evaluation:**

    The performance of the content-based news recommendation system built into this project can be evaluated using various evaluation techniques. The system's effectiveness can be measured using metrics such as **precision, recall, and F1-score**, which can provide insights into the system's ability to recommend relevant articles. With evaluation, we can identify areas of improvement and refine the recommendation algorithms to provide even more relevant recommendations to users.

2. **Collecting Real-Time Data**:

    Currently, the system uses a pre-collected dataset of news articles, but collecting news data in real-time can ensure that the system is always up-to-date with the latest news. An approach that we explored is to use the **Inshorts API** to collect news articles. By combining this with the recommendation system, we can create a more comprehensive news recommendation system that provides real-time recommendations based on the latest news articles. This can improve the relevance and usefulness of the system, as users can receive recommendations based on the most recent news articles available.

3. **Diversity Check**:

   A diversity check mechanism can be implemented to check if the articles being recommended to the user are diverse enough. A diversity score can be calculated and if the diversity score falls below a certain threshold, a report with some statistics about the user's reading behaviour can be generated and the user will be given the option to break out of their filter bubble if they want to.

4. **Serendipity Model:**

   We also want to allow users to break out of their filter bubbles and explore different content. To do this, we can implement a serendipity model, which recommends random articles to users who choose to break out of their usual content. This can help users discover new topics and perspectives they might not have encountered otherwise.

Conclusion:

- Our study examined the effectiveness of a content-based recommendation system for news articles based on their attributes.
- The system used two popular content-based recommendation techniques: Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF).
- The goal of the recommendation system was to recommend news articles similar to a given article based on their attributes, such as article headline, category, author, and publishing date.
- Our study showed that BoW and TF-IDF techniques could effectively recommend news articles based on their attributes. The BoW technique is simple and effective, while the TF-IDF technique considers the frequency of words and produces more relevant recommendations.
- However, both techniques have their limitations and do not capture the semantic and syntactic similarity of words. Word embedding techniques such as Word2Vec, GloVe, and fastText can capture the semantic similarity between words and address the limitations of BoW and TF-IDF techniques.
- The choice of recommendation technique should be based on the specific requirements and characteristics of the data and application domain.

References

1. News recommender system: a review of recent progress, challenges, and opportunities | SpringerLink
2. [2106.08934] Personalized News Recommendation: Methods and Challenges
3. Fairness in Recommendation: A Survey
4. News recommender system: a review of recent progress, challenges, and opportunities | SpringerLink
5. A Serendipity Model for News Recommendation | SpringerLink
6. Personalized News Recommendation Based on Click Behavior
7. FairMatch: A Graph-based Approach for Improving Aggregate Diversity in Recommender Systems
8. Diversification in session-based news recommender systems | SpringerLink
9. Rank and relevance in novelty and diversity metrics for recommender systems