Mid-Sem Project Report

Team Name : Team Gryffindor
Project Name : Content-Based News Recommendation System
School of Engineering and Applied Science, Ahmedabad University
Machine Learning - CSP523, Winter 2021

1st Jeet Shah
AU1841006
Ahmedabad University
Ahmedabad, India
jeet.s3@ahduni.edu.in

2nd Manav Vagrecha
AU1841022
Ahmedabad University
Ahmedabad, India
manavkumar.v@ahduni.edu.in

3rd Parth N. Patel
AU1841028
Ahmedabad University
Ahmedabad, India
parht.p1@ahduni.edu.in

4th Shreyansh Shah
AU1841046
Ahmedabad University
Ahmedabad, India
shreyansh.s@ahduni.edu.in

***Abstract*—This report includes the work progress till the Mid. Semester regarding the project on Content-Based News Recommendation System.**

***Index Terms*—News Recommendation System, Content-Based Filtering,**

I. Introduction & Background

A recommender system is a system that suggests products, services, information, to users based on analysis of data collected by different processes including user interactivity, explicit user ratings, similarity amongst users, similarity amongst items, etc. Thus, there are, in detail, 6 different types of recommendation systems. [3]

- Content-Based : A content-based recommender learns a profile of the new user's interests based on the features present, in objects the user has rated. It's basically a keyword specific recommender system here keywords are used to describe the items. Thus, in a content-based recommender system the algorithms used are such that it recommends users similar items that the user has liked in the past or is examining currently.
- Collaborative : Collaborative recommender systems aggregate ratings or recommendations of objects, recognize commonalities between the users on the basis of their ratings, and generate new recommendations based on inter-user comparisons. Major drawback of Collaborative filtering is based on an assumption that people who agreed in the past will agree in the future and that they will like similar kind of objects as they liked in the past.
- Demographic : This system majorly categorize the users based on attributes and make recommendations based on demographic classes.
- Utility-based : Utility based recommender system makes suggestions based on computation of the utility of each object for the user.
- Knowledge-based / Similarity-based : Knowledge based recommendation works on functional knowledge: they have knowledge about how a particular item meets a particular user need, and can therefore reason about the relationship between a need and a possible recommendation.
- Hybrid : It combines the strengths of more than two Recommender system and also eliminates any weakness which exist when only one recommender system is used.
  - Weighted Hybrid Recommender
  - Switching Hybrid Recommender
  - Mixed Hybrid Recommender

In this work, we are majorly focusing on Content-based filtering. It basically works on user feedback of different articles. There are 2 ways to get user preferences.

- Implicit - Track Click based Activity of user
- Explicit - Requires the user to evaluate viewed articles

Based on the user's preferences build on his/hers interactivity, we will classify all the articles available into 2 categories:

- Relevant to the user
- Irrelevant to the user

Update user history after fixed number of clicks and retrain the model accordingly.

II. Literature Survey

We explored various articles on recommendation systems. After a lot of research we finalised the article ”Using Content-Based Filtering for Recommendation” [3]. It clearly guides through the entire process of content-based recommendation starting from data cleaning to feature extraction to determining cosine similarity. It bifurcates all the available news articles into two classifications, namely - Relevant to the user and Irrelevant to the user.

We also went through various datasets to use in our project including bbc, newsAPI, etc. We finally decided to use Microsoft News Dataset. This dataset is the best suite for our application as it consists of the user activity(history) of 1 million users. User history is a mandatory part for content-based recommendation.

III. Implementation

A. Data Information

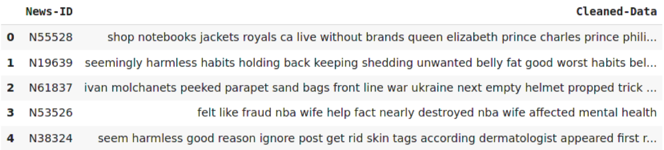
1) Description: MIcrosoft News Dataset (MIND) is a large-scale dataset for news recommendation research. It was collected from anonymized behavior logs of Microsoft News website. The mission of MIND is to serve as a benchmark dataset for news recommendation and facilitate the research in news recommendation and recommender systems area.

2) Resources: Following are few resources from where we can get the dataset

- https://www.kaggle.com/arashnic/mind-news-dataset
- https://msnews.github.io/#about-mind

B. Data Preprocessing

- Here, initially, we removed Redundant Columns from the news and behavior dataset. We preserved important fields like News-ID, Category, Sub-Category, Title, Abstract in the news Dataset and Index, User-ID, History, Impressions in the behaviour dataset.
- Then we removed the rows containing the NaN values for the Title and Abstract fields from the news dataset.
- Then we removed punctuation, stop words and html words - We have used Natuaral Language Toolkit (NLTK) library for removing stop words. [Stop words are the words which occur a large number of times and don't hold major meaning eg. I, the, a, as,...]



	News-ID	Cleaned-Data
0	N55528	shop notebooks jackets royals ca live without brands queen elizabeth prince charles prince phili...
1	N19639	seemingly harmless habits holding back keeping shedding unwanted belly fat good worst habits bel...
2	N61837	ivan molchanets peeked parapet sand bags front line war ukraine next empty helmet propped trick ...
3	N53526	felt like fraud nba wife help fact nearly destroyed nba wife affected mental health
4	N38324	seem harmless good reason ignore post get rid skin tags according dermatologist appeared first r...

Fig. 1. Data after performing Preprocessing

C. Feature Extraction

After obtaining the preprocessed data, we apply N-grams model to it which helps us to obtain unigrams, bigrams and trigrams of the for each article. Here, we will be

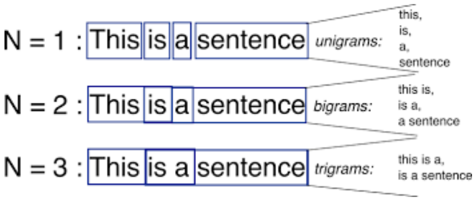
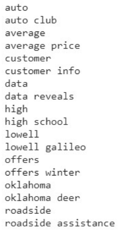


Fig. 2. N-Grams Language Model

Assigning weight to a given feature for a given article that signifies its importance in the given article using TF-IDF.

$$w_i = TF_i \times IDF_i$$



```
auto
auto club
average
average price
customer
customer info
data
data reveals
high
high school
lowell
lowell galileo
offers
offers winter
oklahoma
oklahoma deer
roadside
roadside assistance
```

Fig. 3. Feature Extraction

$$w_i = TF_i \times log(\frac{n}{df_i})$$

where w_i are the weightages for ith term, tf_i = number of occurrences of term t_i in article D, n = number of articles in the dataset, df_i = number of articles in which term t_i appears atleast once

D. Results till date

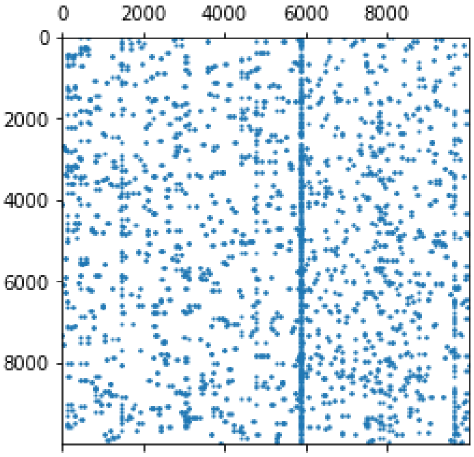


Fig. 4. Visualizing the Feature Matrix

We have got individual feature vector for each article where each feature is assigned a weight as per its significance in the article. Above image shows a cropped portion of the feature matrix whose actual dimensions are 48616x1869854, where 48616 is the total number of articles and 186

IV. Conclusion

This work majorly defines the process of content-based filtering in news recommendation system using MIND MIrcosoft News Dataset. Uptill now, we have implemented data cleaning, data preprocessing and feature-extraction We have arrived to the step of finding the similarity amongst the feature vectors.

References

[1] "Classifying Different Types of Recommender Systems," BluePi, 14-Nov-2015. [Online]. Available: https://www.bluepiit.com/blog/classifying-recommender-systems/. [Accessed: 17-Mar-2021].

[2] "Introduction to MIND and MIND-small datasets," [Online]. Available: https://github.com/msnews/msnews.github.io/blob/master/assets/doc/introduction.md

[3] “Using Content-Based Filtering for Recommendation,” R. van Meteren and M. van Someren, [Online]. Available: http://users.ics.forth.gr/~potamias/mlnia/paper_6.pdf. [Accessed: 17-Mar-2021].