News Recommendation System Report

SABUDH

Prepared By:

Shubham Sharma

Harjeevan Singh

Anchit

Carol Eunice

Ram Mohan Reddy

Abstract

Nowadays there are millions of websites offering news updates from various organisations all over the world. These resources provide valuable information and different perspectives on a specific subject, event or public figure. Personalised recommendation of news and articles is the new way to view our daily topics of interest. The aim of this project is to investigate and build a unique recommender system that can be implemented by news providers in an easy way. Initially, the project introduces the detailed analysis of existing techniques in recommender engines, then focuses on the system design of the engine and the work carried out on the system. The project concludes with the experiments conducted on the engine.

Project Title: Personalised News Recommendation

Engine Author: Shubham Sharma, Harjeevan Singh , Anchit,Carol Eunice, Ram Mohan Reddy

Keywords: Recommender Systems, Collaborative Filtering, Content-based Filtering, Hybrid Recommender Systems, User Modelling

About Assignment

- We as the data scientists are assigned a task of building a news recommendation system by a startup called JhakaasNewsVala.

  Objective:

  → Increase click through rate and frequency of opening the app by the user

  → Reduction in popularity bias

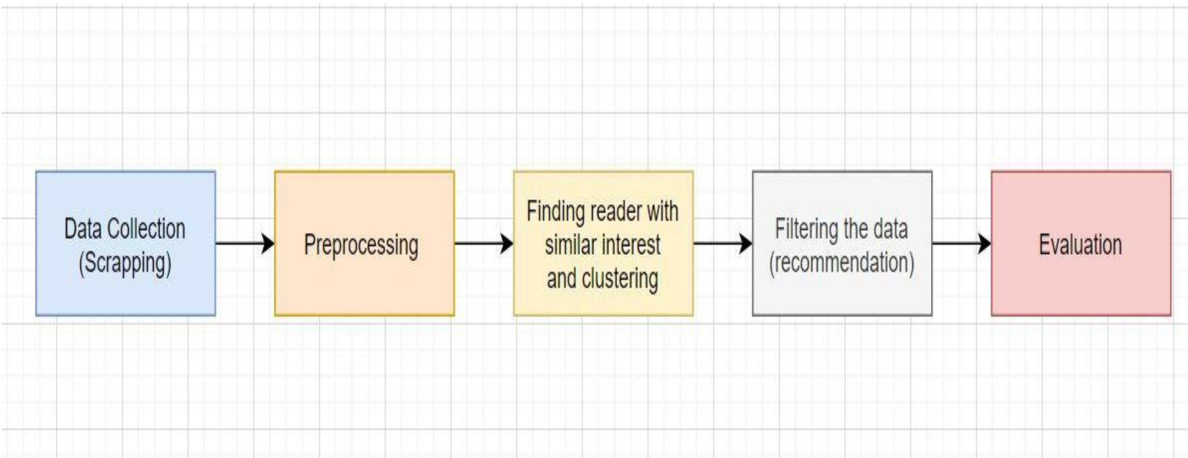
Target audience

Our Target customers

- Working Professionals
- Age Group (21-40)
- Who uses app for reading news

APPROACH

- Data Collection
- Preprocessing
- Tf-Idf vectorizer
- Modelling (Hybrid filtering approach)
- Evaluation



<u>**Datasets**</u>

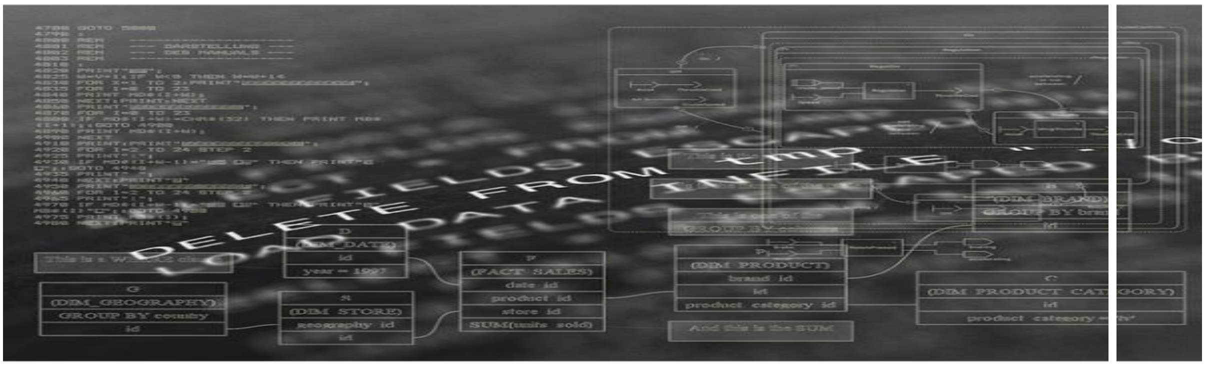
News Articles-
Scraped for India, US and world news from sites
<u>indiatoday.com</u>, <u>deccanchronicle.com</u> etc.
Total <u>10,000</u> news articles.

	A	B	C	D	E	F
1	Id	Title	Summary	Date	Link	
2	1	Locks^ chains: Coronavirus puts Indonesia's mentally ill back in shackles	Padlocks^ shackles and chains are used to cover for a lac	07-Oct-20	https://www.indiatoday.in/world/asia/071020/locks-chains-coronavirus-puts-indonesias-mentally-ill-back-in-shack.html	
3	2	Opposition in Kyrgyzstan claim power after storming government buildings	Kyrgyzstan is a close ally of Russia and has long been a pl	06-Oct-20	https://www.indiatoday.in/world/asia/061020/opposition-in-kyrgyzstan-claim-power-after-storming-government-buildin.html	
4	3	India^ US^ Australia^ Japan to discuss China's growing power in Quad talks	The talks follow recent tensions between China and India	06-Oct-20	https://www.indiatoday.in/world/asia/061020/india-us-australia-japan-to-discuss-chinas-growing-power-in-quad-t.html	
5	4	Taiwan says military under pressure from China as missions mount	China^ which claims democratic Taiwan as its own territ	06-Oct-20	https://www.indiatoday.in/world/asia/061020/taiwan-says-military-under-pressure-from-china-as-missions-mount.html	
6	5	Armenia^ Azerbaijan clashes resume over separatist region	The fighting erupted September 27 and has killed dozens	05-Oct-20	https://www.indiatoday.in/world/asia/051020/armenia-azerbaijan-clashes-resume-over-separatist-region.html	
7	6	Interpol issues 'red notice' for fugitive Thai Red Bull heir over hit-and-run	The charges against Vorayuth^ grandson of Red Bull's co-	05-Oct-20	https://www.indiatoday.in/world/asia/051020/interpol-issues-red-notice-for-fugitive-thai-red-bull-heir-over-hit.html	
8	7	Taiwan scrambles jets for second day as Chinese fighter jets buzz island	Taiwan President Tsai Ing-wen pledged deeper ties with t	19-Sep-20	https://www.indiatoday.in/world/asia/190920/taiwan-scrambles-jets-for-second-day-as-chinese-fighter-jets-buzz-isla.html	
9	8	South Korea to fine church for causing countryÃƒÂ€Â™s largest virus cluster	A fresh wave of infections erupted at a church whose me	18-Sep-20	https://www.indiatoday.in/world/asia/180920/south-korea-to-fine-church-for-causing-countrys-largest-virus-clust.html	
10	9	China begins military drills amid US envoy's second high-level visit to Taiwan	Concerned over the ever-closer relationship between Tai	18-Sep-20	https://www.indiatoday.in/world/asia/180920/china-begins-military-drills-amid-us-envoys-second-high-level-visit-t.html	
11	10	Shinzo Abe's entire cabinet resigns as Suga set to become Japan's new PM	Yoshihide Suga^ the new leader of the Liberal Democrati	16-Sep-20	https://www.indiatoday.in/world/asia/160920/shinzo-abes-entire-cabinet-resigns-as-suga-set-to-become-japans-new.html	
12	11	Yoshihide Suga manages an easy win to be Japan's ruling party leader	Given the LDP's legislative majority^ Suga is expected to	14-Sep-20	https://www.indiatoday.in/world/asia/140920/yoshihide-suga-manages-an-easy-win-to-be-japans-ruling-party-leader.html	
13	12	China brands Hong Kong citizens held at sea 'separatists'	The Shenzhen city police said the 12 Hongkongers were u	14-Sep-20	https://www.indiatoday.in/world/asia/140920/china-brands-hong-kong-citizens-held-at-sea-separatists.html	
14	13	South Korea to temporary ease virus curbs in Seoul ahead of festival	Chuseok is one of the South Korea's biggest holidays with	13-Sep-20	https://www.indiatoday.in/world/asia/130920/south-korea-to-temporary-ease-virus-curbs-in-seoul-ahead-of-festival.html	
15	14	Thailand's plane cafes are helping customers pretend they are in the sky	Hungry diners appear even to have missed plane food as	13-Sep-20	https://www.indiatoday.in/world/asia/130920/thailands-plane-cafes-are-helping-customers-pretend-they-are-in-the-s.html	

Code Snippet

```
for page in range(1, pagesToGet+1):
    print('processing page :', page)
    url = 'https://www.deccanchronicle.com/world/americas?pg=' + str(page)
    print(url)

    # an exception might be thrown, so the code should be in a try-except block
    try:
        # use the browser to get the url. This is suspicious command that might blow up.
        page = requests.get(url)  # this might throw an exception if something goes wrong.

    except Exception as e:  # this describes what to do if an exception is thrown
        error_type, error_obj, error_info = sys.exc_info()  # get the exception information
        print('ERROR FOR LINK:', url)  # print the link that cause the problem
        print(error_type, 'Line:', error_info.tb_lineno)  # print error info and line that threw the exception
        continue  # ignore this page. Abandon this and go back.
    time.sleep(2)
    soup = BeautifulSoup(page.text, 'html.parser')
    frame = []
    links = soup.find_all('div', attrs={'class': 'col-sm-12 SunChNewListing'})
    print(len(links))

    for j in links:

        news_id = str(sr_no)
```

Recommendation System



In today's digital world, a recommendation engine is one of the most powerful tools for marketing. A recommender system is nothing but an information filtering system composed of machine learning algorithms that predict a given customer's ratings or preferences for an item. A recommendation engine helps to address the challenge of information overload in the e-commerce space. Thus, it can help in saving a lot of browsing time for customers, as the recommendation engine directs the user to products of he is most likely to like. Its personalization features improve customer engagement and retention. The idea of recommendation engines is also something you are already familiar with; Whether it is product recommendations on Amazon, movie recommendations on Netflix, or music suggestions on YouTube, recommender systems are already supporting many aspects of your experience online.

Types of recommender systems

Broadly based on their operations recommendation engines can be divided into 3 types:

Collaborative filtering : Focuses on analyzing customer behavior, activities or preferences in order to predict ratings or suggest products. Collects large amounts of information on customers' behavior, activities or preferences in order to predict what users will like based on the similarity with other users. Customer attributes like demographics and psychographics are used in identifying similar customers. Amazon is the pioneer in implementing collaborative filtering; it works on collecting preferences from distinct users from which a customer * product matrix is developed. As we see in the following figure user (3) and user(m-1) have similar likes, so we can recommend item(n) to user(3)

	$item_1$	$item_2$	$item_3$	...	$item_n$
$user_1$		5	2		1
$user_2$	3				
$user_3$	1		3		
. . .					
$user_{m-1}$	5		4		2
$user_m$		4			3

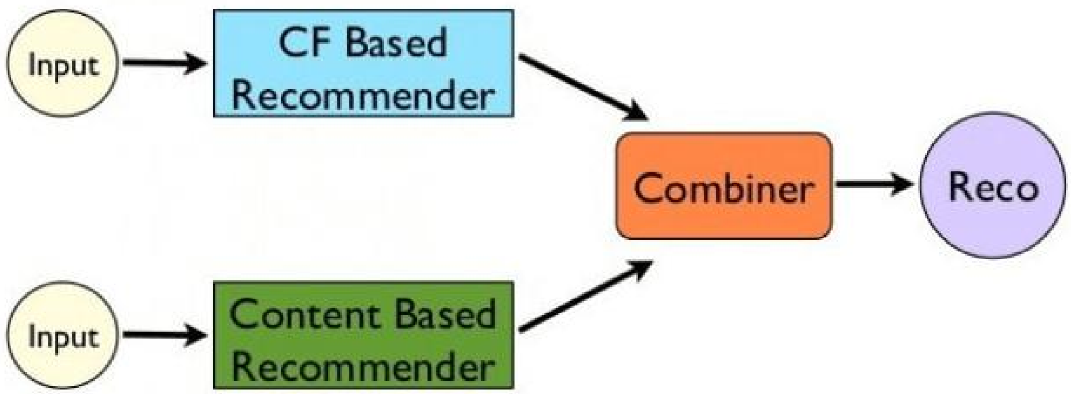
Collaborative filtering is further divided into user-item and item-item. User-item filtering looks for like-minded customers based on their common rating patterns. In item-item filtering similarity between pairs of items is calculated. To summarize, collaborative filtering works on the principle: you are likely to like what others similar to you like ..... . Techniques like matrix factorization are used in collaborative filtering.

Content Based System: The core idea of content-based filtering is: "if you like an item you will also like a 'similar' item.... ". Algorithm recommends products that are similar to the past transactions. Similarity of the items or the neighborhood is computed with techniques like cosine and euclidean distances. A item-feature matrix is created by computing the tfidf values from product descriptions.

Major differences between content based and collaborative filtering

	Content-Based Filtering	Collaborative Filtering
User Properties	The *types* of things the user has **purchased, watched, ranked, clicked**	The User's purchase patterns
	EXAMPLE: User X has purchased a book that is: • over 30K words • written by author Y • has a red cover • Sci-fi genre	EXAMPLE: User X has purchased/watched, etc. • Item 1 • Item 2 • Item 3 • and so on...
Item Properties	Properties of the item	A list of users that have bought the item
	EXAMPLE: The book is: • over 30K words • written by author Y • has a red cover • in the Sci-fi genre	EXAMPLE: Item Y has been purchased/watched/etc by: • User 1 • User 2 • User 3 • and so on...

Hybrid model approach : Leverage both item metadata and transaction data to give recommendations. It combines the content-based and collaborative- based models. After evaluating the performance of pure recommendation engines (content & collaborative based) and hybrid model, it is observed that the hybrid model outperforms. Netflix is a good example of the hybrid model implementation. It takes into account features of movies along with interest of users. Here, using natural language processing (NLP), tags can be generated for the movie based on its story, and then tfidf scores can be used to calculate the similarity between the products and collaborative filtering can be used to recommend movies to the user depending on their features.



Challenges of developing a recommender system

Due to the recommender system insufficient data suffers from both the cold-start and sparsity problem. Cold start in general refers to the difficulty to instantiate the recommender system.

Product cold start and user cold start are two distinct cold start issues. Product cold start occurs when a new product is launched it lacks valuable user interactions, thus the engine fails to target the right group of customers. The product cold start issue can be addressed through content-based filtering- the metadata of the new product can be used to compute its similarity with already existing products.

User cold start challenge arises when a customer visits the engine for the very first time. The recommender fails to direct the customer to the best possible options since there is no past behavior monitored to understand his likes/dislikes or preferences. Suggesting most popular products aligning to the search can lead to some customer activity. Data sparsity arises when users in general interact with limited number of products from the available potential products. Clustering similar users and products together can be one of the feasible solutions to address sparsity.

Our Thinking Pipeline

Step 1: Started with Content Based Approach

We consider first content based approach as it is basically the technique in which we don't need data about other users because this model's recommendations are specific to the current user. This makes it easier to scale to a large number of users.
But we feel that this model has very limited ability to expand on user's existing interests.
So we then inclined towards collaborative filtering technique.

Step2: Inclination towards Collaborative Filtering

In collaborative technique we think about Cold Start Problem. The cold start problem is the major problem in all recommendation systems based on collaborative filtering.
The problem raises when the new user joins the system and doesn't have any clicks. There is no data about the user to recommend items.
This problem is an obvious case when the system is initiated for use or when the system has high item-user ratio.
It is more prominent in news domain because new user visits after an event has occurred or users who occasionally visits news apps based on expected news articles to be published.

Step 3 : Solution to cold start problem

The way in which we can solve visitor cold start problem is by providing them the most popular articles overall or regionally. Product cold start problem can be solved by using content based approach as it is less prone to popularity bias.

Popularity Bias: It is a condition in which handful of items get high interactions and most of them only receive a fraction of them.
Since content based recommendation choose which item to recommend based on the feature the item possess, even if no interaction for a new item exist, still its feature will allow for a recommendation to be made.

So we finalize on using content based approach to tackle this problem but without depending upon user based features like reviews and tags.

<u>Feedback Problem</u>

Explicit Feedback -> The explicit feedback of user plays an important role in precise recommendation of that new article to the same news reader. Explicit feedback can be in the form of comments,click on like/dislike,click on sharing feature.
Problem : Like everyone is not so interested in giving feedback so we come up with implicit feedback technique.

Implicit Feedback -> The system should be able to conceive implicit feedback from the news reader for effective recommendation and the privacy of the user should keep intact.
Form of Implicit Feedback :

a) Clicks on an article (Generated using binomial and exponential distribution)
b) Time spend of reading an article (Generated using GMM)

Now we wanted to built a system which will give equal importance to user interests and also to user's non-interesting factor .
So how to get user's non- interesting factor-
By user's click on dislike feature in our app
By sentiment analysis of comments section of that user.
By time spent by the user below a threshold.

If we have recommended like **k articles** to user and he leaves **top p article** and click on **p+1 article** then we got to know that our active user <u>don't like</u> **top p articles** out of k articles so that we can leverage this fact to generate user preferences.

****