

Mini Project Report on

PERSONALIZED NEWS RECOMMENDATION SYSTEM

Submitted in partial fulfillment of the requirement for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE & ENGINEERING

Submitted by:

Student Name

Harshit Dutt Tyagi

University Roll No.

2018825

Under the Mentorship of

Mr. Arnav Kotiyal

Assistant Professor



**Department of Computer Science and Engineering
Graphic Era (Deemed to be University)
Dehradun, Uttarakhand
July 2024**

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the project report entitled **“Personalized News Recommendation System”** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineering of the Graphic Era (Deemed to be University), Dehradun shall be carried out by the under the mentorship of **Mr Arnav Kotiyal, Assistant Professor**, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Harshit Dutt Tyagi

2018825


A handwritten signature in black ink that reads 'Harshit'.

Table of Contents

Chapter No.	Description	Page No.
Chapter 1	Introduction	4
Chapter 2	Literature Survey	6
Chapter 3	Methodology	8
Chapter 4	Result and Discussion	11
Chapter 5	Conclusion and Future Work	12
	References	13

Chapter 1

Introduction

In the following sections, a brief introduction and the problem statement for the work has been included.

1. Introduction

Traditional news consumption has shifted from physical newspaper subscriptions to online platforms. Aggregators like Google News and Yahoo! News collect news from various sources, presenting a global view of news. However, the sheer volume of articles can overwhelm users, necessitating effective filtering mechanisms. Personalized news recommendation systems play a crucial role in helping users discover relevant news articles amidst the overwhelming volume of available content.

Personalized news recommendation systems have become increasingly important in the digital age. These systems aim to help users discover news articles that align with their interests while mitigating information overload.

For users who are logged in and have enabled web history, personalized recommendation systems build profiles based on their past click behavior. These profiles capture users' news interests over time, allowing the system to understand how their preferences evolve. Large-scale analysis of anonymized click logs helps predict users' current news interests based on their activities and overall news trends.

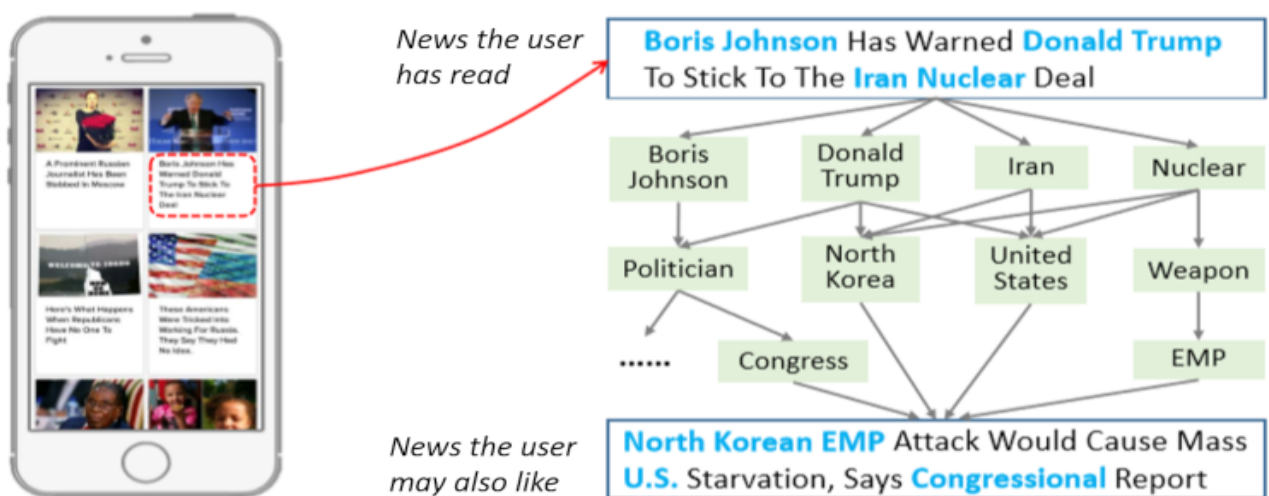


Figure 1.1 Example of Personalized News Recommendation [1]

Upon implementing personalized news recommendation systems, we observe Improved User Engagement. Users receive relevant articles aligned with their interests, leading to higher engagement. Personalization reduces information overload and enhances the likelihood of users returning to the platform.

Personalized news recommendation systems have become indispensable in the digital age, where an overwhelming amount of information floods users daily. Traditional newspaper subscriptions have given way to online news platforms, necessitating efficient ways to filter and present relevant content. In this context, personalized recommendations play a pivotal role.

The primary goal of personalized news recommendation systems is to tailor news articles to individual users' preferences, interests, and browsing behavior. By doing so, these systems enhance user engagement, increase click-through rates, and ultimately improve the overall user experience. As we delve deeper into the topic, we'll explore various techniques and methodologies employed in achieving this personalization.

Chapter 2

Literature Survey

A comprehensive literature survey reveals a rich landscape of research and development in personalized news recommendation systems. Researchers have explored diverse approaches, including:

- 1. Collaborative Filtering:** Collaborative filtering leverages user-item interaction data to make recommendations. It identifies patterns by analysing users' historical behaviour (such as clicks, likes, and shares) and identifies similar users or items. Matrix factorization techniques, such as Singular Value Decomposition (SVD) and Alternating Least Squares (ALS), are commonly used in collaborative filtering. Challenges include the cold-start problem (for new users or items) and scalability issues.
- 2. Content-Based Filtering:** Content-based filtering focuses on the intrinsic characteristics of news articles. It considers features like article text, keywords, and metadata. By building user profiles based on their interactions with specific content, content-based methods recommend articles similar to those the user has previously engaged with. However, content-based approaches may struggle with serendipity (introducing users to novel content) and diversity.

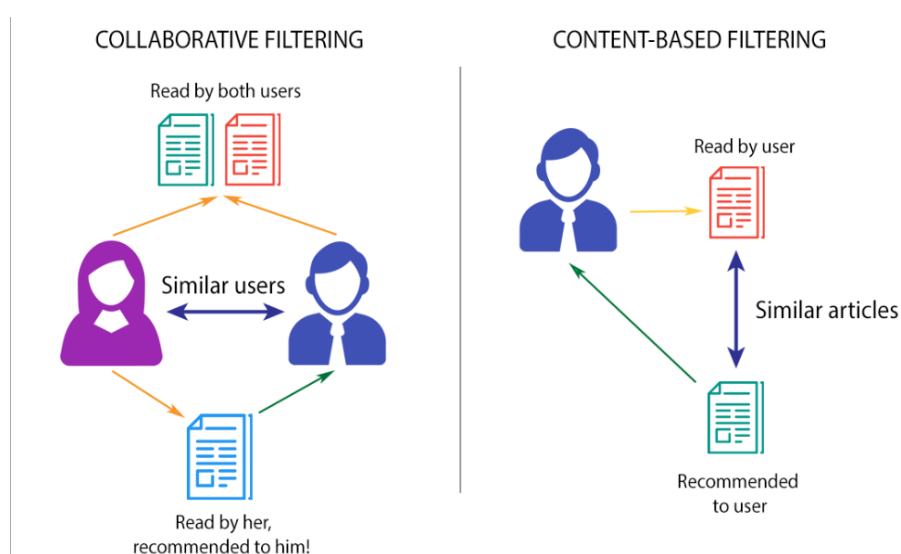


Figure 2.1 Comparing types of filtering approaches

3. **Hybrid Approaches:** Hybrid models combine collaborative filtering and content-based techniques to mitigate the limitations of each. For instance, hybrid systems can use collaborative filtering to address the cold-start problem and content-based filtering to enhance recommendation diversity.

Xiangfu Meng and his team[2] also categorized news recommendations in following types:

1. **Time-Based News Recommendation:** Time-based news recommendation focuses on the timeliness and recency of news articles. News articles have a short lifespan, and their relevance diminishes over time. Approaches include using decaying weights for older articles or considering only recent news. The challenge lies in recommending relevant news while accounting for its temporal context.
2. **Location-Based News Recommendation:** Location-aware news recommendation considers users' geographical contexts. Recommendations are based on the user's location, such as recommending nearby restaurants, events, or points of interest. Researchers explore two main directions: physical distance-based recommendations, which consider proximity, and geographical topic-based recommendations, which focus on location-related topics.
3. **Social Networks-Based News Recommendation:** Social networks play a crucial role in information dissemination. Users share news and interact with content within their social circles. Recommendations leverage social relationships, user profiles, and shared information. Social network graphs guide personalized news suggestions.
4. **Session-Based News Recommendation:** Session-based recommendation focuses on short-term user intentions. Sessions encapsulate articles a user interacts with in a brief time period. Combines content-based and collaborative filtering to address data sparsity. Adapts quickly to changing user interests.
5. **Multi-Modal News Recommendation:** Multi-modal approaches incorporate both textual and visual information from news articles. News representations learn from news texts and images. Images provide additional context and attraction for users. Techniques include encoding news texts and image regions-of-interest (ROIs) using pre-trained models

Chapter 3

Methodology

We decided to develop a content-based prediction model over a web-scraped dataset. The result will be displayed on a website using Streamlit.

3.1 LIBRARIES

The following libraries were used in the program.

1. **NumPy**: Offers various mathematical functions to operate on data structures.
2. **Pandas**: Enables data cleaning, transformation, and exploration over DataFrames, a tabular data structure similar to a spreadsheet.
3. **Natural Language Toolkit**: Provides tools for tokenization, stemming, tagging, parsing, and semantic reasoning of human language data (text).
4. **Sklearn**: It is a machine learning library. Provides tools for feature extraction, model evaluation, and hyperparameter tuning.
5. **Beautiful Soup**: Parses HTML and XML documents.
6. **Selenium**: used for web Scraping dynamic web pages.
7. **Pickle**: Allows saving objects to files or transferring them over networks.
8. **Streamlit**: It is a framework for creating web apps using Python.

```
import numpy as np
import pandas as pd
import nltk
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics.pairwise import cosine_similarity
from bs4 import BeautifulSoup
import selenium
import streamlit as st
import pickle
import streamlit.components.v1 as components
```

Figure 3.1 figure showing every library imported for the program

3.2 DATA COLLECTION AND PREPROCESSING

web scraping and preprocessing are essential steps. Web scraping allows you to gather relevant data from various websites. Preprocessing ensures data consistency, accuracy, and reliability.

Web Scraping

Web scraping refers to the process of extracting data from websites or web pages. Python along with libraries like BeautifulSoup allows you to parse HTML and extract relevant information from web pages.

Steps:

- Retrieve HTML: Use the `urlopen` function to fetch the HTML content of a web page.
- Create a BeautifulSoup Object: Parse the HTML using BeautifulSoup to convert it into Python objects.
- Extract Data: Identify relevant tags, classes, or elements within the HTML structure and extract the desired data.
- Clean and Process: Preprocess the extracted data (e.g., remove duplicates, handle missing values).
- Store Data: Save the cleaned data in a suitable format (e.g., CSV, JSON) for further analysis.

Preprocessing

Data preprocessing involves cleaning, transforming, and organizing raw data to prepare it for analysis or modeling.

Steps:

- Handling Missing Values: Impute or remove missing data points.
- Data Transformation: Normalize, scale, or encode features.
- Feature Engineering: Create new features or derive meaningful information.
- Removing Outliers: Identify and handle extreme values.
- Text Preprocessing: Tokenization, stemming, and removing stop words for textual data.
- Data Integration: Combine data from multiple sources.
- Data Reduction: Reduce dimensionality (e.g., PCA) if needed.

3.2 PREDICTION

Once our data is cleaned, structured, and ready for further exploration, modeling, or visualization, we decided on using a content-based filtering model.

Count Vectorizer

The CountVectorizer is a powerful tool provided by the scikit-learn library in Python. It plays a crucial role in transforming text data (such as news articles) into numerical vectors based on each word's frequency (count). Given a collection of news articles, the CountVectorizer creates a matrix where each unique word (or term) is represented as a column. Each news article corresponds to a row in this matrix. The value in each cell represents the count of a specific word in that particular article. The representation shown above is known as a sparse matrix. Words not present in an article have a count of 0, resulting in a matrix with many zeros.

Cosine Similarity

Cosine similarity measures the similarity between two vectors in a multi-dimensional space. It calculates the cosine of the angle between the vectors, indicating how closely they align. Values range from -1 (completely dissimilar) to 1 (identical), with 0 indicating orthogonality (no similarity). Represent each news article as a vector in a high-dimensional space. The dimensions correspond to features such as keywords, topics, or other relevant attributes. To recommend news articles, compute the cosine similarity between pairs of article vectors. Higher cosine similarity indicates greater similarity between articles.

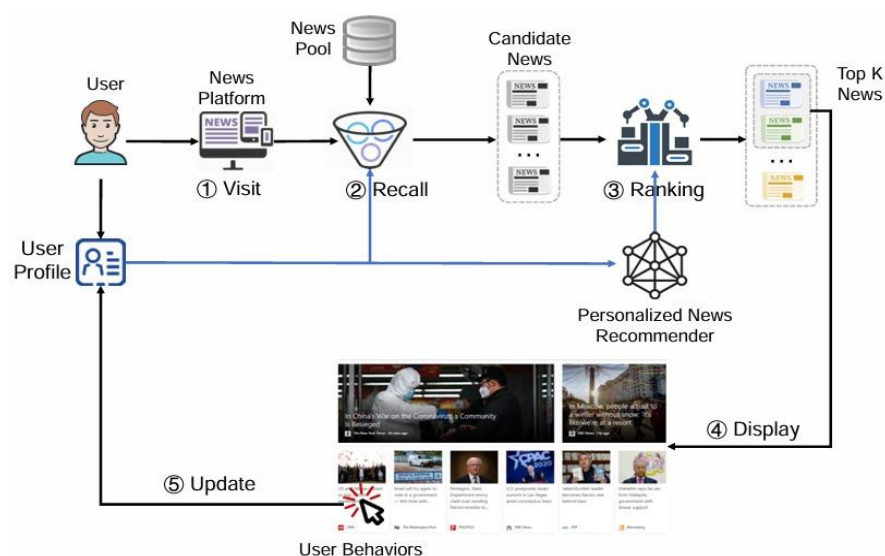


Figure 3.2 Figure showing the prediction process[3].

Chapter 4

Result and Discussion

The user selects one of the news articles collected from the web [4][5]. Then we map the similarity matrix of that article with the dataset and sort it based on the highest similarity. Then the 5 most similar articles are displayed as recommendations. Further, we can use streamlit to show the result on a web page.

```
I Was An NBA Wife. Here's How It Affected My Mental Health.

RECOMMENDATIONS:
Health problems treated by acupuncture

How To Reach Out If Your Friend Is Struggling With Their Mental Health

Injuries Across the NBA

Ranking all 30 NBA head coaches

Women's mental health improves after giving up alcohol, study finds
```

Figure 4.1 figure showing the Result

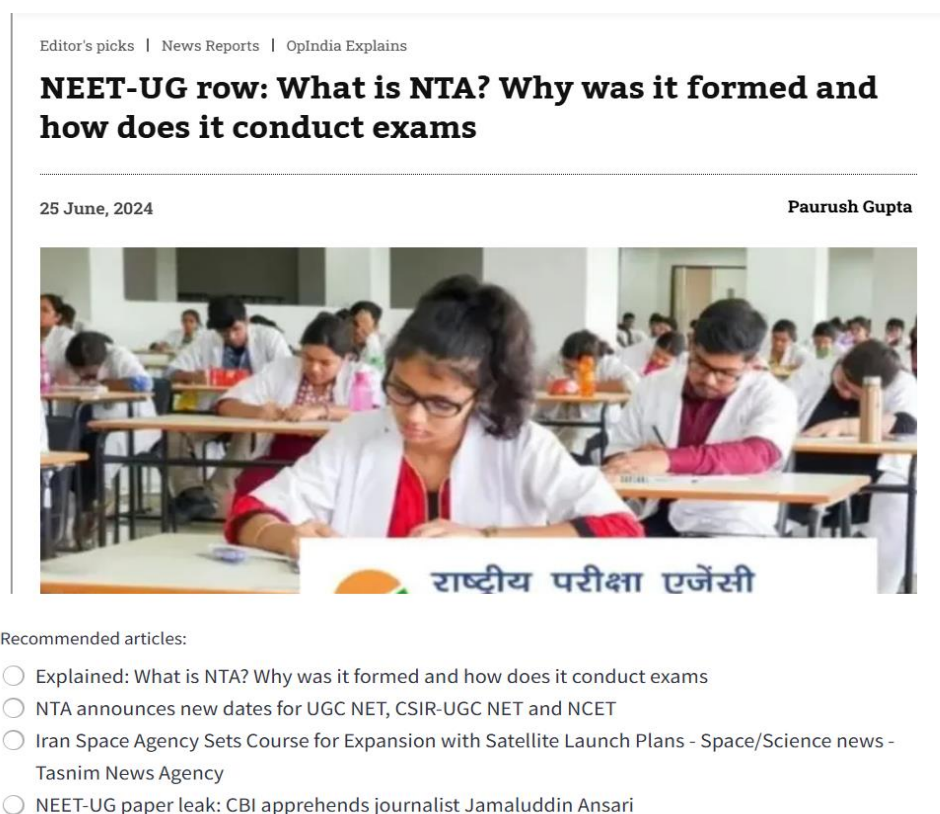


Figure 4.2 figure showing the web page of the Result

Chapter 5

Conclusion and Future Work

In conclusion, personalized news recommendation systems are pivotal in today's information-rich landscape. As research continues addressing challenges and refining algorithms will lead to even more effective and user-friendly systems. These systems empower users to stay informed while navigating the vast sea of news content.

Further, a more hybrid approach can be used in combination with stored sessions for each user login. More complex deep-learning models can be used for mapping user interactions. We can also add knowledge graphs, reinforcement learning and user profiling for better results.

References

- [1] [Personalized Recommendation Systems: Five Hot Research Topics You Must Know - Microsoft Research](#)
- [2] A Survey of Personalized News Recommendation Xiangfu Meng¹ · Hongjin Huo¹ · Xiaoyan Zhang¹ · Wanchun Wang¹ · Jinxia Zhu¹ Received: 14 April 2023 / Revised: 13 July 2023 / Accepted: 16 August 2023 / Published online: 2 September 2023 © The Author(s) 2023
- [3] Chuhanwu, Fangzhaowu, Yongfenghuang, Xing Xie, arXiv:2106.08934v3 [cs.IR] 24 Feb 2022
- [4] [OpIndia - Latest India News, Updates, Analysis, Opinions, Social Media news](#)
- [5] [VOA - Voice of America English News \(voanews.com\)](#)