Chapter 3
A Lightweight Content-Based News Recommendation System Using TF-IDF and Cosine Similarity

Snehal Rahul Rathi
https://orcid.org/0009-0004-7638-6429
Vishwakarma Institute of Technology, Pune, India

Aditi Meshram
Vishwakarma Institute of Information Technology, Pune, India

Ritik Narote
Vishwakarma Institute of Information Technology, Pune, India

Shilpa Kalantri
https://orcid.org/0000-0002-7011-2513
Shah and Anchor Kutchhi Engineering College, Mumbai, India

ABSTRACT

The exponential growth in digital media has revolutionized news consumption. However, it has also led to significant challenges in filtering relevant content from the vast information pool. Personalized news recommendation systems are essential to ensure user engagement and satisfaction. This paper proposes a machine learning-based News Recommendation System leveraging Term Frequency-Inverse Document Frequency (TF-IDF) and cosine similarity to identify and suggest similar

DOI: 10.4018/979-8-3373-3063-1.ch003

news articles based on content alone. The system is implemented using Python and Streamlit with additional features like category filtering, keyword search, and article export. Evaluation on a dataset of 1000+ real-world news article demonstrates the model's effectiveness and potential for scalable deployment.

INTRODUCTION

The growth of online content has completely changed how we get news. In the past, most people read newspapers or watched TV to stay informed. These days, we can use apps on our phones or visit websites to get news quickly. This means that anyone can access news at any time. However, having so much information available can make it difficult for people to find news that suits their personal interests. Generic news platforms often present the same trending articles to every user, disregarding their preferences, location, or past reading behavior. Recommendation systems can solve this by tailoring content suggestions. However, many existing systems rely heavily on user history, which introduces privacy.

This paper presents a content-based News Recommendation System designed for real-time use, built using TF-IDF vectorization and cosine similarity. It also includes practical features such as article preview, similarity scoring, and the ability to filter by category and keyword – all delivered via a responsive Streamlit web interface.

In contrast, content-based systems require no personal data and offer comparable accuracy in certain domain.

The system is implemented using Python 3.10 and several powerful open-source libraries, ensuring ease of use and reproducibility.

BACKGROUND

In the contemporary world, people are inundated with an excess of news because of the internet. The overwhelming content available on multiple platforms makes it harder for the reader to find articles that piques their interest. Regular news websites provide trending and standard articles, which do not cater to the individual user's needs, as there is no algorithmic personalization to discover content. To enhance user engagement and recommend pertinent articles, an advanced and smart system is needed. This task is best approached using content-based filtering techniques which recommend items based on their inherent properties and aim to provide similar items to the user, especially when user history is sparse or nonexistent.

The recommendation system we designed uses content-based filtering techniques with Term Frequency-Inverse Document Frequency (TF-IDF). This model consid-