

Profile-Based Retrieval

Information Retrieval, Extraction & Integration

Course 2022/23

Víctor Morcuende Castell

Guillermo Nájera Lavid

TABLE OF CONTENTS

[1. Introduction 2](#_Toc131633103)

[2. Problem Context and Dataset Selection 4](#_Toc131633104)

1. Introduction

2. Problem and dataset

3. Preprocessing and cleaning

4. Profile's Creation

5. Encoding of the documents

6. Profile-Based Retrieval Implementation

7. Performance assessment

8. Results

9. Conclusions

# **Introduction**

The proliferation of digital content on the internet has led to an ever-increasing volume of information available to users. While this provides access to a vast array of knowledge, it also poses the challenge of efficiently finding relevant and personalized content. In response to this challenge, profile-based information retrieval systems have emerged as an effective solution for tailoring content recommendations to individual users based on their unique interests and preferences.

Profile-based information retrieval systems construct user profiles that encapsulate users’ interests, preferences, and previous interactions with content. These profiles serve as the basis for personalized content recommendations, enabling the system to present users with content that is more relevant to their individual preferences. Such systems have widespread applications across various domains, including news websites, e-commerce platforms, and entertainment services.

The primary objective in this assignment is to design, implement, and evaluate a content-based recommendation system that identifies and recommends articles tailored to user profiles. The system aims to accurately discern articles that are relevant to users’ interests, resulting in a more personalized and engaging user experience. In pursuit of this goal, we will explore different techniques for constructing user profiles, such as the “Weighted Topic Frequency” (WTF) method and assess the efficacy of various classifiers in pinpointing relevant content.

Our project will comprise several stages, beginning with data preprocessing and feature extraction, followed by user profile creation using the WTF method. Subsequently, we will employ similarity measurements to establish connections between user profiles and documents, leading to personalized content recommendations. Finally, we will evaluate the system’s performance using a range of metrics and assess the results obtained from multiple classifiers.

In the forthcoming sections, we will delve into the intricacies of our content-based recommendation system, providing a comprehensive overview of each stage in the process. The data preprocessing and feature extraction phases will be outlined, followed by a detailed explanation of user profile creation using the WTF method. Next, we will discuss the method for measuring similarity between profiles and documents, which forms the foundation for generating content recommendations. The system evaluation will then be presented, encompassing the various performance metrics employed, as well as an in-depth analysis of the results obtained from different classifiers. Through this extensive exploration, we aim to provide a thorough understanding of our profile-based information retrieval system and its potential for delivering highly personalized content recommendations.

# **Problem Context and Dataset Selection**

The problem we are addressing in this project is the efficient and accurate recommendation of content tailored to individual users. With the vast amount of information available on the internet, it is crucial to provide users with relevant and personalized content to enhance their overall experience. By doing so, we can help users find valuable information more efficiently and encourage their continued engagement with the platform.

To tackle this problem, we require a dataset that contains a substantial number of articles with sufficient information to construct meaningful user profiles and develop a robust recommendation system. For this purpose, we have chosen the “BBC News Train” dataset, which is a collection of news articles from the BBC. The dataset comprises 1.490 articles spanning across five different categories: business, entertainment, politics, sport, and tech.

The choice of the BBC News Train dataset offers several advantages. Firstly, the diverse range of topics covered in the dataset allows us to create user profiles that encompass a variety of interests, making it suitable for developing a versatile recommendation system. Secondly, the dataset contains high-quality and well-structured articles, ensuring that our system can accurately capture the nuances of the content. Additionally, the dataset has been used in numerous research studies and machine learning competitions, which facilitates the comparison of our results with existing work in the field.

In summary, the problem we aim to address is the accurate recommendation of personalized content to users based on their unique profiles. The “BBC News Train” dataset, with its extensive collection of articles spanning a diverse array of topics, provides an ideal resource for developing and evaluating our profile-based information retrieval system.

# **Preprocessing and Cleaning the Dataset**

To develop an effective profile-based information retrieval system, it is essential to preprocess and clean the dataset to ensure that it is in a suitable format for analysis. This process involves several steps, which are described in detail below.

1. Importing the dataset: The first step is to import the dataset into our Python environment. We read the CSV file containing the BBC News Train dataset and store it in a suitable data structure, such as a Pandas DataFrame, for easy manipulation and analysis. This allows us to access and process the data in an efficient and intuitive manner.
2. Removing duplicate articles: To ensure that our dataset is accurate and representative, we must remove any duplicate articles that may be present. Duplicate articles can lead to biased results and affect the performance of our recommendation system. We achieve this by identifying and removing rows with identical values in the “Text” column using the ‘drop\_duplicates()’ function available in Pandas. This helps us maintain the integrity of the dataset and prevent any distortions in the results of our analysis.
3. Text cleaning: As the dataset contains textual data, it is crucial to clean and preprocess the text to ensure it is in a suitable format for analysis. This process typically involves several steps:
   1. Converting the text to lowercase: This helps ensure that the text is consistent and that words are not treated as separate entities due to differences in case.
   2. Removing punctuation marks: Punctuation marks can introduce noise and make it difficult to process the text. By removing them, we simplify the text and facilitate analysis.
   3. Removing stop words: Commonly used words such as 'and', 'the', and 'in' do not carry much meaning and can be safely excluded from the analysis. This reduces the size of the dataset and improves the efficiency of subsequent processing steps.
   4. Tokenization: Breaking the text into individual words, or tokens, enables us to analyze the text at the word level and identify meaningful patterns and relationships.
   5. Lemmatization: this technique reduces words to their root forms, further simplifying the text and reducing the number of unique words in the dataset. This can improve the performance and efficiency of the analysis.
4. Creating a clean dataset: After completing the preprocessing and cleaning steps, we are left with a clean dataset that is ready for further analysis. This dataset should be free of duplicates, missing values, and irrelevant information, and the textual data should be preprocessed and simplified to ensure it is in a suitable format for analysis. With this clean dataset, we can proceed to develop our profile-based information retrieval system, extract relevant features, and evaluate its performance using appropriate metrics and evaluation techniques.

# **Profile’s Creation**

The creation of user profiles is a pivotal component in the profile-based information retrieval system. This section delves deeper into balancing the dataset, the reasoning behind selecting the TF-IDF method for vectorizing documents and profiles, and the formation of topics and profiles.

**Balancing the dataset**

In the context of a multi-class classification problem, it is crucial to maintain a balanced dataset to prevent potential biases in the model's performance. These biases could lead to overemphasis on certain categories and neglect of others, thereby compromising the model's accuracy and efficiency. To achieve a balanced dataset, the code provided ensures an equal number of articles are present for each category. This approach guarantees that the model is exposed to a representative sample of each topic, which in turn enhances its ability to discern between different categories and provide more accurate results.

## **TF-IDF for vectorizing documents and profiles**

The Term Frequency-Inverse Document Frequency (TF-IDF) method was chosen for vectorizing the documents and user profiles due to its robust and widely accepted performance in text representation. TF-IDF's popularity stems from its ability to convert text into numerical format, a requirement for machine learning algorithms to effectively process and analyze data. The TF-IDF technique calculates the importance of each term within a document based on its frequency in the document and its rarity across the entire corpus. Consequently, it helps identify terms that are significant and distinctive, which can then be utilized as features to differentiate between various topics.

## **Topics and profiles**

The selected dataset comprises news articles from five distinct categories: business, entertainment, politics, sport, and tech. These categories were chosen to represent a wide range of topics, which allows for testing the model's efficacy in accurately retrieving relevant information based on user profiles. To create the user profiles, a set of keywords was identified for each category. These keywords encapsulate the category's primary themes and act as the foundation for defining user interests. The user profiles were then generated by merging these keywords with their corresponding TF-IDF weights, resulting in a numerical representation of the user's interests.

## **Profile’s creation process**

The code provided initially computes the TF-IDF matrix for the entire dataset, which serves as the base for profile creation. For each category, a list of representative keywords is defined, and their corresponding TF-IDF values are extracted from the matrix. These values are then averaged to create a centroid for each category, representing a typical user interested in that category. This process leads to the formation of five distinct user profiles, one for each category, which can be employed to evaluate the performance of the profile-based information retrieval system.

The meticulous creation of user profiles is a vital step in the development of an effective profile-based information retrieval system. By ensuring a balanced dataset, utilizing the TF-IDF method for vectorization, and carefully crafting topics and profiles, the model is poised to achieve optimal results in retrieving relevant information tailored to user preferences and interests.