INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 4 - V7I4-1188)
Available online at: https://www.ijariit.com

# Effectively developing a recommendation system by implementing collaborative filtering

*Darshan C. M. Pawar*
pawarsupreeth@gmail.com
*Atria Institute of Technology, Bengaluru, Karnataka*

*Arthi Sharma V.*
arthisharma13@gmail.com
*Atria Institute of Technology, Bengaluru, Karnataka*

*Syeda Fathima Zohra*
sfathimazohra@gmail.com
*Atria Institute of Technology, Bengaluru, Karnataka*

*Syeda Arbeena*
arbeena469@gmail.com
*Atria Institute of Technology, Bengaluru, Karnataka*

*Maqdum Shariff*
maqdum.shariff@atria.edu
*Atria Institute of Technology, Bengaluru, Karnataka*

**ABSTRACT**

***Recommender structures are developed to predict a client's preferences and recommend items that are likely to be relevant to them. They are most likely the most complicated AI computations used by web businesses to assist with transactions. Collaborative filtering, for example, finds a group of users based on the goods they buy or provide comments on, and then recommends popular items in the group. Using variables like Video ID, hate, likes, favourite count, description, and keyword, a video recommendation engine is employed on the YouTube dataset. The most popular online video community on the planet is YouTube. Users' likes and dislikes on the site are used to suggest groups of videos to them. Collaborative content filtering algorithms were used in the suggested system. Data can be collected in a variety of ways, such as downloading with certain categories, ensuring that the information is always up to date. Users will see the top five YouTube videos based on the experimental results.***

***Keywords—*** *Recommendation System, Machine Learning, Collaborative filtering, Python.*

## 1. INTRODUCTION

YouTube is one of the most famous online video community which has huge amount of user generated video content. Youtube is an american video platform where users may create, share, and watch videos... Due to availability of large number of videos on YouTube, it is necessary to have a good recommendation system. Recommendation system plays a vital role in identifying user's interest and recommending the videos that user may like. The more tailored the recommendations are to the user's interests, the longer the preference based on the evaluations of other users with similar interests. In recommender systems, there are two major approaches:1) Collaborative filtering is a method of predicting a user's suggests the user's interest in a given item, 2) A data recommendation user stays on the site and views the recommended videos because the provided item description and the user profile are comparable, or because the given item and things picked by the user are similar.

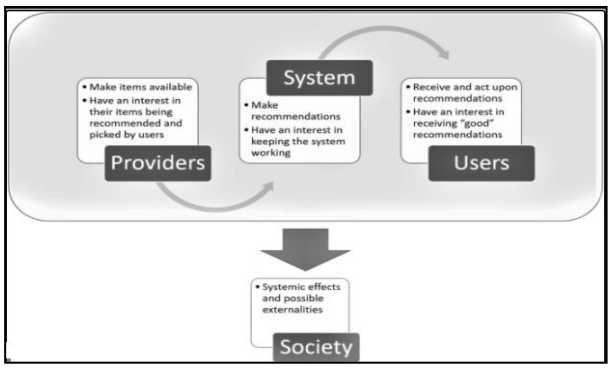


**Fig. 1: Recommendation Overview**
***(Source: Gooogle.com)***

### 1.1 Data collection

The selection of high-quality data for analysis is part of the data collection process. We used a dataset from youtube.com in this study. User can assign few keywords, for those keywords, we can download real time dataset from youtube. An information investigator's job is to find new means and sources for obtaining significant and far-reaching data, interpreting it, and analysing the results using factual processes.

### 1.2 Data Pre-processing

Pre-handling is motivated by the need to transform raw data into an AI-friendly design. A data scientist can receive more accurate findings from an AI model when the data is organised and tidy. The method combines data planning, cleansing, and evaluation.

 

**Non-functional requirements**

The list of non-functional requirements is provided below. Internal stakeholders will need to define the particular details.

- Time to Respond
- Accessibility
- Regularity
- Maintainability
- Accessibility

### 1.3 Python

Python is a comprehensively beneficial programming language that can be translated to a verified level. Python, which was created by Guido Van Rosum and initially released in 1991, has an arrangement hypothesis that stresses code clarity and the use of fundamental whitespace. It offers fabricates that allow for straightforward programming on both small and large scales. Python has a fantastic sort structure and memory leaders that've been changed. It supports many programming ideal models, such as object-oriented, fundamental, utilitarian, and procedural, and includes a large and comprehensive standard library. python interpreter's are available for a large number of operating systems. The reference version of Python, CPython, as well as it is free and open source software with a community-based development approach in virtually all of its variation applications. The python Software Foundation, a non-profit organisation, is in charge of CPython.

## 2. SYSTEM MODEL

Context planning is the process of putting up the engineering, segments, modules, interfaces, and information for a framework to satisfy specified models. System planning is the application of frameworks hypothesis to item improvement.

### 2.1 System Architectural Design

The structure and behaviour of a system are defined by a conceptual model called system architecture. It includes the system components as well as the relationships that describe how they interact to create the overall system. The Fig 2 below shows the system's architecture and the various components added to them. The architecture design, the dataset for implementation, the method employed, and the UML designs are all covered in this chapter. The above figure represents system architecture of proposed system, where we are applying Collaborative filtering to get results.

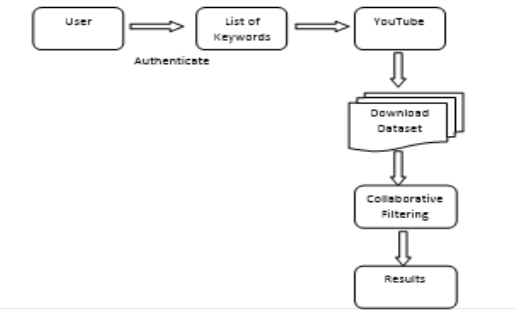


**Fig 2: System Architecture**

### 2.2 Use Case Diagram

A use case graphic shows a user's involvement with the system at its most basic level by demonstrating the relationship between the user and the numerous use cases in which the client is engaged. A use case diagram can be used to indicate the many categories of client of a system as well as the several use-cases, and it is frequently supplemented by other diagrams. While a use case may go into great depth about each option, a usecase diagram can help provide a higher-level understanding of the system. " The blueprints for your system are use case diagrams." someone once stated. They give a simpler and graphical picture of what the system needs to achieve. The self-explanatory use-case diagram is as shown in the following fig 3.

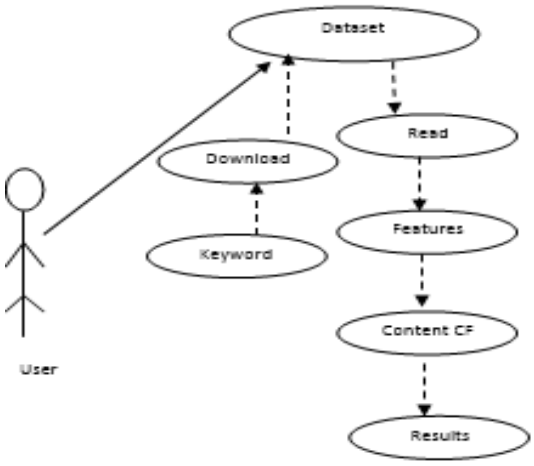


**Fig 3: Use Case Diagram**

## 3. IMPLEMENTATION METHODOLOGY

System Implementation uses the structure developed during architectural design and the outcomes of system analysis to produce framework for elements that fulfil stakeholder and criterion for the system set in the early life cycle. These framework components are then combined to provide middle-of-the-road totals, which are subsequently combined to form the overall arrangement of-interest (SoI). Execution is the process of producing the lowest level framework elements in the framework pecking order. System components are created, obtained, or reconditioned. The shaping, removing, connecting, and finishing of hardware, the writing and testing of software, and the creation of operating procedures for operators are all part of the hardware manufacturing process responsibilities all fall under the category of production. Modular design, often known as "modularity in design," is a technique for breaking down a system into smaller pieces called modules or skids that may be created independently and then reused in other systems. The functional segmentation of a modular system into discrete, scalable, and reusable modules, as well as the strict usage of well-defined modular interfaces and adoption of industry standards for interfaces, characterise a modular system. The work described here is built in Python 3.6.4 and makes use of scikit-learn, pandas, matplotlib, and other necessary tools.We downloaded real-time datasets from YouTube, including ratings and users. A data mining algorithm is one of them. The Collaborative Filtering technique is used.

### 3.1 Dataset Collection and Pre-Process

Researchers can utilise the dataset to test recommender systems and collaborative filtering techniques. It could be used as a testbed for matrix and graph techniques such as PCA and clustering. Data pre-processing is now a data mining approach that entails converting unstructured data into a usable format. Data from the real world is frequently partial, inconsistent, and deficient in specific behaviours or trends, and it is likely to contain several inaccuracies. Pre-processing information is an attempted strategy for dealing with such issues. User can assign few keywords, for those keywords, we can download real time dataset from youtube. The secure authentication key for user is mandatory to download dataset from live data. Videoid, comments counts, dislike counts, like counts, favorite count, title, number of views, description are the keyword.

 

| commentCou | dislikeCount | favoriteCount | likeCount | videoId | title | viewCount | description | keyword |
|---|---|---|---|---|---|---|---|---|
| 186 | 1248 | 0 | 5777 | DIlWAdmmQzM | Funny Pranks Lol JUST FOR LAUGH | 1821020 | Funny Pranks Lol JUST FOR LAUGH GAGS 201 | laugh |
| 1744 | 3414 | 0 | 16490 | vQToLSbBwbw | Try Not to Laugh 27 Best Funny Fa | 5668817 | Buy T-Shirt & More Now! https://lifeawesoi | laugh |
| 359 | 1859 | 0 | 15148 | ODyI7avvW9Ek | â–¶ BEST 2019 Just For Laughs Gag | 2603913 | Click Link Subscribe: https://goo.gl/kDPY2A | laugh |
| 245 | 458 | 0 | 3404 | 9_7SkwBqtLY | TRY NOT TO LAUGH - EPIC FAILS V | 879390 | Brand new weekly fail compilation of the fi | laugh |
| 653 | 3595 | 0 | 35213 | bgqm3KjEq34 | Top 5 Just For Laughs Gags - July 2 | 9186317 | Hello July 2018! Just For Laughs has release | laugh |
| 104 | 203 | 0 | 1384 | QSMacmunsI0 | TRY NOT TO LAUGH - BEST FAILS V | 281032 | Brand new weekly fail compilation of the fi | laugh |
| 540 | 6862 | 0 | 23720 | NCJF3B2PDJk | Just For Laugh 2019 ðŸ˜‚Ø±Ø¨Ø | 12399021 | Music use in video: Itro & Tobu - Cloud 9 [Ni | laugh |
| 473 | 4970 | 0 | 22793 | NsKurka2gKw | BEST Just For Laughs Gags Prank ! | 6395116 | BEST Just For Laughs Gags Prank ! Best Cand | laugh |
| 484 | 4143 | 0 | 18120 | YfbBFoHc7m4 | Try Not to Laugh Challenge - Top | 7986518 | Buy T-Shirt & More Now! https://lifeawesoi | laugh |
| 17009 | 10246 | 0 | 209374 | rSbx8SGsPc | TOP 10 Funniest Comedians That | 22127087 | Watch the funniest and down right hilariou | laugh |
| | 4384 | 0 | 23396 | c4IpEyAWtf0 | Best Epic WATER FAILS â"... Try Nc | 7966583 | Best Epic WATER FAILS â"¯ Try Not To Laugh ! | laugh |
| | 7985 | 0 | 61750 | 2t2ZrgtkGfs | TRY NOT TO LAUGH CHALLENGE | | 3699170 For new episode every Thursday : SUBSCRIE | laugh |
| 1113 | 7125 | 0 | 33068 | EDvvFCrrkJQ | Try Not To Laugh Funny Fails 2018 | 11969164 | Buy T-Shirt & More Now! https://lifeawesoi | laugh |
| 38 | 9 | 0 | 140 | MWceNlQa-wY | Hilarious Awesome Pics To Make | 1341 | Funny Pictures That Will Make You Laugh av | laugh |
| 22902 | 1838 | 0 | 182801 | qnsaSb6cgIs | L A U G H | 3445861 | LAUGH! Support the channel: https://youtu | laugh |
| 2320 | 4461 | 0 | 45721 | DkXlUNCN6VM | Just For Laughs Gags - Best Off | 19118011 | I sell on EBAY wordpress ,PHP,Android,java | laugh |
| 270 | 1361 | 0 | 10569 | fBUg70UKS4 | Top 5 Just For Laughs Gags - Nove | 2515859 | Hello November 2018! It's nearing the end c | laugh |
| 120 | 4 | 0 | 390 | S6IPFsVGEpY | PETER GOES SKYDIVING! \| Family | 2758 | I react to Family Guy Funniest Moments. Pe | laugh |
| 5222 | 10056 | 0 | 120067 | NH2WBcg1kOg | Try not to laugh - Random video | 8528189 | Support this channel at https://www.patrec | laugh |
| 8985 | 7904 | 0 | 87820 | abDb0ImYLiE | Try Not To Laugh Watching Funny | 14857733 | Try Not To Laugh Watching the best Kids Fai | laugh |
| 2907 | 13054 | 0 | 69510 | I7j6FP8Oam4 | Try Not to Laugh Funny Kids Fails | 16413795 | Try Not to Laugh Funny Kids Fails Jan 2019 \| | laugh |
| 62 | 6 | 0 | 505 | 2gxgO2craG0 | Renegades React to... Markiplier | 9156 | WANT EARLY ACCESS AND UNEDITED REACT | laugh |

**Fig 4: Pre-Processed Dataset.**

### 3.2 Collaborative-Filtering Algorithm

Collaborative filtering is a way for sifting through items that a client could appreciate based on the replies of like clients. It works by sifting through a large group of people and identifying a smaller group of clients with similar likes to a given client. It looks at the items they enjoy and compiles them into a prioritised list of suggestions. There are a variety of methods for determining which clients are similar and combining their decisions to create a list of proposals. This post will show you how to do it the best manner possible with Python. Collaborative Filtering technique of filtering or evaluating objects utilising the opinions of others is known as collaborative filtering. While the term "collaborative filtering" has only been around for a few years, it is based on something that humans have done for centuries - sharing thoughts with others. Show a user a list of objects in order of their potential utility. This is sometimes defined as anticipating the user's rating of an item and then ranking the things based on that projected rating.

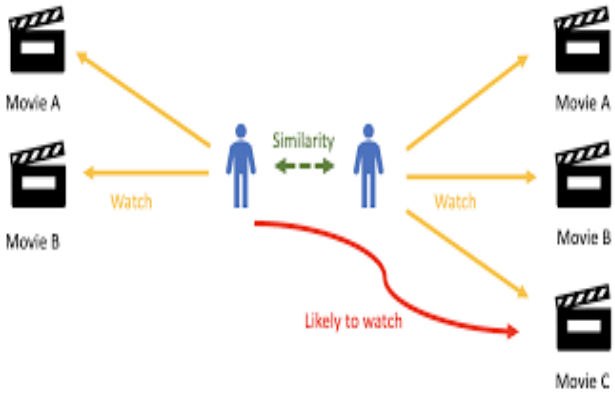


**Fig 5: Activity Diagram**

**Algorithm for Content Based Collaborative Filtering**

**Step 1:** Examines the collection of things rated by the target user, calculates how identical they are to the target item, and then chooses the k most similar items.

**Step 2:** The forecast is calculated using a weighted average of the rating of the target user on the most comparable items.

**Step 3:** The likeness between items I and j is calculated by separating the people who rated them and then using a similarity computation approach.

**Step 4:** Cosine-based Similarity — in the m-dimensional user space, objects are vectors.

### 3.3 Scikit-learn

Scikit-learn is a set of easy-to-use tools for data mining and analysis. Scikit-learn is a Python-based machine learning package that is available for free. It includes support vector machine, random forrests, gradient boosting, k-means, and DBSCAN, among other classification, regression, and clustering techniques, and NumPy and SciPy, Python's numerical and scientific libraries, have been built to operate with it.

Some popular groups of models provided by scikit-learn include:

- Clustering is a technique for organising unlabeled data, such as KMeans.
- Cross Validation is a technique for assessing the performance of supervised models using data that has never been seen before.
- Datasets: for testing and generating datasets with specific characteristics to investigate model behaviour.
- Dimension Reduction: used to reduce the number of characteristics in data for purposes of outline, representation, and highlight selection, such as Principal Part Investigation.
- Ensemble methods are used to combine the predictions of many administered models.
- Extraction of features is used to characterize credits in image and text data.
- Optional features: for identifying important traits from which to build guided models.
- Parameter tuning is used to make the most of directed models.
- Manifold Learning: For condensing and visualizing complex multi-dimensional data.
- Supervised Models: a broad category that includes summing direct models, separate analysis, naive bayes, lazy techniques, neural networks, support vector machines, and decision trees, among others.

## 4. TESTING

### 4.1 Introduction

The next difficult and time-consuming phase after completing the creation of any computer-based system is system testing. Only the development company knows how far the user requirements have been satisfied during the testing process, and so on. Software testing is an essential part of software quality assurance since it is the last check of the specification, design, and code. The expense of software failures, as well as the rising practicality of software as a system, are driving pressures for thorough testing.

- **Testing Objectives:** There are a few rules that may be used as testing objectives:
  - Testing is the process of running a programme with the goal of detecting a mistake.
  - A excellent test case is one that has a high chance of uncovering a mistake that has yet to be identified.
- **Source Code Testing:** This analyzes the system's logic. We may state that the reasoning is flawless if we are obtaining the output that the user expects.

### 4.2 Specification Testing

We can specify what the software should do and how it should behave in certain situations. This testing is a side-by-side comparison of system performance and requirements over time.

### 4.3 Module Level Testing

The errror's will be identified at each individual module, encouraging the programmer to find and fix faults without affecting other modules.

### 4.4 Unit Testing

Unit testing is concerned with ensuring that the smallest unit of software has completed its task. The integrity of the date saved temporarily is verified against the local data structure during the execution of the algorithm. Boundary conditions are checked to

 

make certain that the module works properly at the limits or restrictions set by the user.

### 4.5 Validation Testing

It starts following the successful completion of the incorporation testing. Approval is successful when the product performs in a way that the client can understand. The majority of the approval occurs during the information flow measure, when there is the greatest risk of supplying incorrect data. Other validation will be carried out in all processes where the proper details and data must be submitted in order to obtain the desired results.

### 4.6 Output Testing

The output of the proposed system must be checked after the validation testing, as no system can be deemed useful until it provides the right output in the specified format. The screen format and the printer format are two types of output formats.

### 4.7 User Acceptance Testing

User Acceptance Testing (UAT) is an important part of any system's success. By keeping in touch with possible users throughout the development process and making adjustments as needed, the system under consideration is put to the test for user approval. Testing is the process of checking all of the test cases for mistakes and correcting them. This method is repeated for each unit and each unit is tested separately.

**Table 1: Test Cases**

| Test Case ID | Test Description | Test Procedure | Test Input | Expected Result | Actual Result |
|---|---|---|---|---|---|
| T101 | To download dataset | Click on Create Dataset button | Execute Main.py | Dataset to be loaded | Alert to "Dataset Downloaded" |
| T102 | To check algorithm function | Click on CF button after without loading dataset | Execute Main.py | Alert should be given for dataset upload | Alert to "Load Dataset" |
| T104 | To check algorithm function | Click on CF button after loading dataset | Execute Main.py | Evaluation metrics to be done | Alert to "CF Successfully Finished" |

## 5. RESULT

YouTube Recommender System (YRS) is proposed and a recommendation as a result of it The YRS is built using data gathered from the YouTube site through their API. The data is made up of many numbers of videos as well as comments on each one. Users are suggested videos based on collaborative filtering, a common data mining technique. The number of videos that can be suggested to viewers is restricted to five. By obtaining the dataset in real time, users may access popular videos. The data may be gathered in a variety of ways, such as downloading with specific categories, which guarantees that the dataset is always up to date. There are some restrictions to download datasets by using the YouTube API. As a result, we've limited the number of videos in each category to twenty. Users are presented with the top five YouTube titles based on the testing findings.


**Fig 6: The application main page**


**Fig 7: User entered query collaborative filtering algorithm works and calculate results**

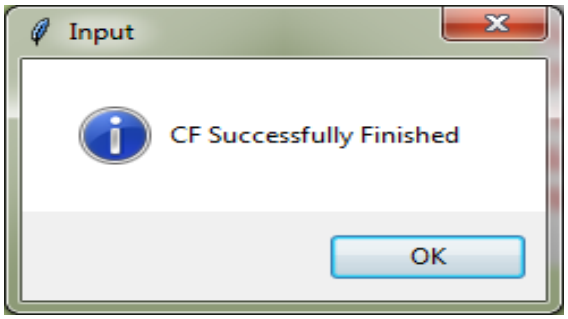


**Fig 8: User gets alert after process completed**


**Fig 9: User gets the top five videos from the dataset results**

## 6. CONCLUSION AND FUTURE ENHANCEMENTS

In the proposed work, YouTube is the most popular online video community in the world, with a large volume of user-generated video material. YouTube is a international video-sharing platform that allows clients to create, share, and watch videos. A suggestion is obtained by the YouTube Recommender System (YRS). The YRS is built using data gathered from the YouTube

 

website through their API. The data is made up of a number of videos as well as comments on each video. Users are suggested videos based on collaborative filtering, a common data mining technique. The number of videos that can be suggested to viewers is restricted to five. By obtaining the dataset in real time, users may access popular videos. There are certain restrictions. The dataset was downloaded using the YouTube API. Thus we have limited the number of videos to twenty in each category. As a future updates, we are spellbound to extends this for some hybrid estimations or potentially deep learning computation to convey with more effective outcome.

## 7. REFERENCES

[1] V. Furtado, A. Melo, A. L. V. Coeho, and R. Menezes, "A crime simulation model based on social networks and swarm intelligance," in Proceedings of the 2007 ACM Symposium on applied Computing (SAC), 2007, pp. 56–57.

[2] P. Arabia & Y. Windi, "Marketing and social networks," in Advances in social network analysis: Research in the social and behavioral science and evelopment, S. Wasserman and J. Galaskiewicz, Eds. Sage Press, 1994, pp. 254–273.

[3] A. S. Klovdahl, "Social networks and the spread of infectious diseases" Social Science and Medicine, vol. 21, no. 11, pp. 1203–1216, 1985.

[4] B. Coingsworth and R. Menezes, "Identification of social tension in organizational networks," in Complex Network, Studies in Computational Intelligance, S. Fortunato, G. Mangioni, R. Menezes, and V. Nicosia, Eds. Springer Verlag, 2009, pp. 209–223.

[5]Googlle Inc., "YouTube Fact Sheet," http://www.youtube.com/t/fact sheet, 2006.

[6] J. Paolilo, "Structure and network in the youtube core," in Hawai International Conference on System Sciences. IEEE Computer Society, 2008, pp. 146–156.

[7] T. Mei, B. Yang, X. Hua, L. Yang, S. Yang, and S. Li, "VideoReach: an online video recommendation system," in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007, pp. 767–768.

 