MOVIE RECOMMENDER SYSTEM

Presented By :
Aswini Purnima Sreepada
Department of Computer Science

Georgia State University

AGENDA

- Introduction
- Motivation & Objective
- Environment Set up
- Data Overview
- Methodology
- Results of Implementation
- Conclusion

INTRODUCTION

Recommender systems attempt to forecast users' interests and propose product items that are likely to be of interest to them. They are among the most advanced algorithms used by online businesses to generate sales and understand client behavior.

INTRODUCTION ..contd

Netflix, Inc. is an American Technology & media-services provider and production company whose Primary business is its subscription-based streaming service which offers online streaming of a library of films and television programs.

As of January 2022, Netflix has about 222 million paid subscriptions worldwide, including 173 million from United States.

This project provides a systematic and intelligent use of methodologies for data analysis using multiple python packages like Numpy, Networkx and pandas_profiling. This involves statistical analysis of the metrics like Mean, Median, Mode, Sum and Standard deviation.

MOTIVATION

Any movie recommendation system gives a degree of comfort and customization that allows the user to connect with the system more effectively and view movies that are relevant to his requirements. The key motive for choosing a movie recommendation system as my Project is to provide this degree of comfort to the user.

The fundamental purpose of the proposed movie recommendation systems is to analyze the user data consumption and forecast which movies any user is most likely to want to view.

OBJECTIVE

- Perform Exploratory Data analysis on the data released on Netflix for 90+ years (1925-2020)

- To improve recommendations services upon the Traditional one on Netflix to make the recommendation system significant and efficient.

- To visualize insights of the exploratory analysis and present working solution of the recommender system.

DATA OVERVIEW

- The dataset consists of meta details about the movies and Television shows such as the title, director, and cast of the shows / movies, the release year, the rating, duration etc.

- For current analysis, consider only the movies and TV shows data released between December 1925 – January 2020.

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	81145628	Movie	Norm of the North: King Sized Adventure	Richard Finn, Tim Maltby	Alan Marriott, Andrew Toth, Brian Dobson, Cole...	United States, India, South Korea, China	2019-09-09	2019	TV-PG	90 min	Children & Family Movies, Comedies	Before planning an awesome wedding for his gra...
1	80117401	Movie	Jandino: Whatever it Takes	NaN	Jandino Asporaat	United Kingdom	2016-09-09	2016	TV-MA	94 min	Stand-Up Comedy	Jandino Asporaat riffs on the challenges of ra...
2	70234439	TV Show	Transformers Prime	NaN	Peter Cullen, Sumalee Montano, Frank Welker, J...	United States	2018-09-08	2013	TV-Y7-FV	1 Season	Kids' TV	With the help of three human allies, the Autob...
3	80058654	TV Show	Transformers: Robots in Disguise	NaN	Will Friedle, Darren Criss, Constance Zimmer, ...	United States	2018-09-08	2016	TV-Y7	1 Season	Kids' TV	When a prison ship crash unleashes hundreds of...
4	80125979	Movie	#realityhigh	Fernando Lebrija	Nesta Cooper, Kate Walsh, John Michael Higgins...	United States	2017-09-08	2017	TV-14	99 min	Comedies	When nerdy high schooler Dani finally attracts...

DATA STATISTICS

Dataset statistics	
Number of variables	12
Number of observations	6234
Missing cells	3036
Missing cells (%)	4.1%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	5.8 MiB
Average record size in memory	973.8 B

METHODOLOGY

Sub - Graph Mining Technique:

The Sub-graph mining algorithm employs a candidate generation technique based on edges. Every iteration increases the size of the sub-graph by one edge (k+1) Patterns are merged only if both share the same k-edged sub graph. Graphs are categorized by the edges in the edge-disjoint path method making the vertices that are connected to P through the Path length <=1 condition.

Algorithm 01 – Sub-Graph Mining Summarized

Function: Sub-graph Mining

Input: G: Graph for Analysis

P: Basic Pattern Candidate

V_{new}: Vertex related to P

Output: Mp: List of patterns (P) occurrences = P ∪ V_{new}

1: Create an empty list Mp

2: **foreach** vertex v in G do

3: Calculate the maximum path length (*l*) between P and v

4: Obtain set of vertices connected to P through a path length <=1

5: Identify the subset V*l* of vertices bearing label ~ V_{new}

6: Add every graph P ∪ {v}, with v ϵV*l* added to new list Mp:

7: End **foreach**

8: Return Mp

METHODOLOGY

The connections and closeness between nodes and edges are established using Adamic Adar measure which computes the closeness of nodes based on their shared neighborhood. This measure is defined as follows:

$$\text{AdamicAdar}\ (x,y) = \sum_{u \in N(x) \cap N(y)} \frac{1}{\log(N(u))}$$

For each node u in common to x and y, add to the above measure, where x and y are two movies/ TV shows. The below observations are noted:

- If x and y share a node u that has a lot of adjacent nodes, this node is not really relevant. → N(u) is high → 1 / log (N(u)) is not high
- If x and y share a node u that not has a lot of adjacent nodes, this node is really relevant. → N(u) is not high 1 / log (N(u)) is higher.

METHODOLOGY ..contd

The workflow around the established graphical data is as follows:

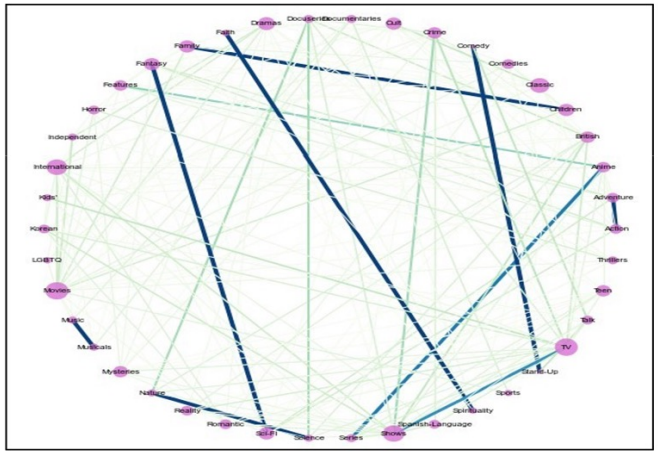
- Explore the neighborhood of the target film → this is a list of actor, director, country, categories;
- Explore the neighborhood of each neighbor → discover the movies that share a node with the target field;
- Calculate Adamic Adar measure → Final Movie Recommendation (3 accurate genre matching movies, total of up to 5 movies)

DEMONSTRATION

Georgia State University

RESULTS OF IMPLEMENTATION

- Input Dataset obtained from: https://data.world › chasewillden › netflix-shows-movies, to create a networkx graph by utilizing the networkx library
- The graph is obtained by considering the attribute “Listed_in” to identify the correlation between each of the values of the attribute, as shown:



A recommendation system is proposed using the Exploratory data analysis of movies and TV Shows.

The recommender system uses Adamic Adar measure to measure the closeness of shared neighborhood of nodes.

Initially, K-means clustering is performed for all the movies in the dataset using TF-IDF (term frequency–inverse document frequency)matrix.

The properties of the undirected graph shown are: Nodes present in the graph -

- Movies
- Person (actor or director)
- Categories
- Countries
- Cluster (description)
- Sim(title) top 5 similar movies

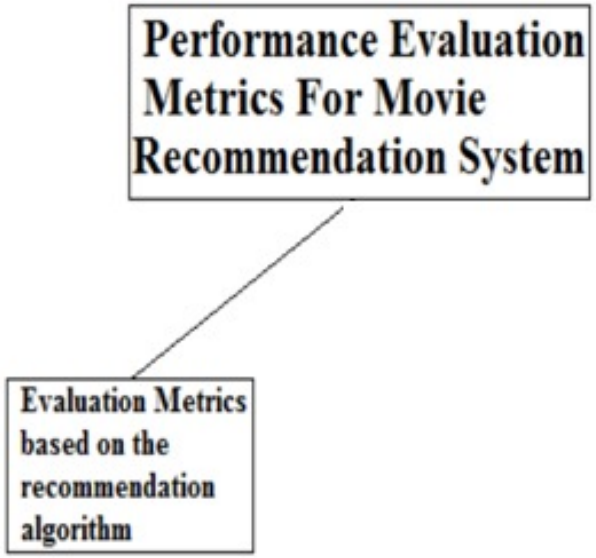
Edges present in the graph –

- ACTED_IN : relation between an actor and a movie
- CAT_IN : relation between a categries and a movie
- DIRECTED : relation between a director and a movie
- COU_IN : relation between a country and a movie
- DESCRIPTION : relation between a cluster and a movie
- SIMILARITY in the sense of the description

Evaluation Metrics

Evaluation Based on Recommendation Algorithm:

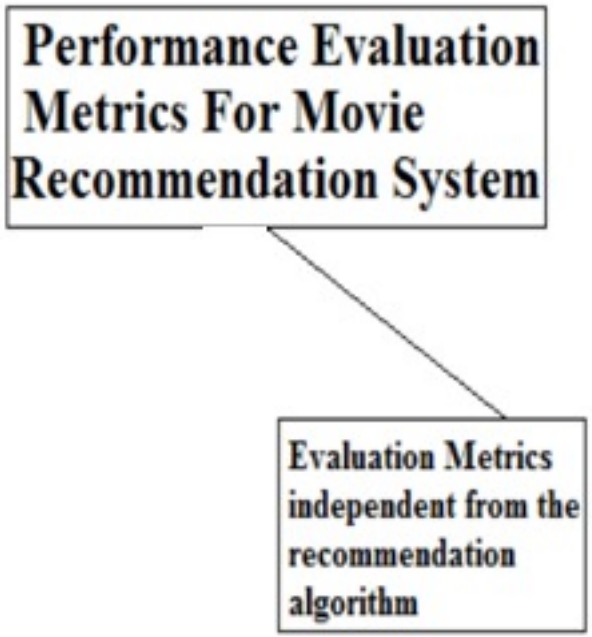
- Diversity
- Novelty
- Serendipity
- Accuracy
- Mean Absolute Error
- Recall
- Precision
- Coverage
  - Prediction Coverage
  - Catalogue Coverage



Evaluation Metrics

Evaluation Metrics Independent of Recommendation Algorithm:

- System Perspective
  - Scalability
  - Confidence
  - Sparsity
  - Adaptivity
  - Robustness: Accuracy, Stability
  - Computing Time: Training phase, Run Time Phase
- User Perspective:
  - User Preference
  - Trust
  - Privacy



Overall Experimentation Results

Most of the movies are classified with at least **2 genres.** The findings include:

→Average rating for most of the movies is 3.07 with 75% of the rating given to movies are within **3.5** rating.

→Average count of rating given to a movie is 423 with very high standard deviation of **2477**

→Drama is the most filmed genre with 8637 films (roughly 1 in 7) as standalone and 14624 (roughly 1 in 4) as one of genre of a movie

→Avg rating given by a user is 3.67 which is also the average rating for the movies

REFLECTIONS

- ✓ Mining common graph patterns takes a long time in general, thus different strategies for exploring relevant sub graphs without creating the whole pattern collection are applied.

- ✓ Our preliminary analysis revealed that many users receive uncalibrated recommendations. These uncalibrated recommendations can be calibrated using sub-graph mining techniques to generate better outcomes.

- ✓ Using the recommender system built using the Adamic Adar measure it is noted that the performance of the current recommender system in place is better than the traditional recommendation system that Netflix has in place.

Questions & Suggestions?

Georgia State University

THANK YOU

Georgia State University