CPS 803 Group 21 Project Proposal: Movie Recommender System Analysis

Amos Pivato
Computer Science
Ryerson University
Toronto, Canada
apivato@ryerson.ca

Haoyu Chen
Computer Science
Ryerson University
Toronto, Canada
haoyu.chen@ryerson.ca

Haoyu Zhang
Computer Science
Ryerson University
Toronto, Canada
haoyu.zhang@ryerson.ca

***Abstract*—Proposal of an application level project using machine learning classification algorithms in a movie Recommender System and comparing results between approaches.**

I. Introduction

Recommender Systems are a class of algorithm and method that can recommend or suggest content that is most likely to be relevant for a specific use case. Typically relevant to a user on how likely they are to interact with a given product. The purpose of this system is to learn and predict which content users will prefer. There are typically two types of Recommender System, content based which compare similarity of items or content and collaborative filtering which compares the interactions between like minded users and the content they have interacted with.

II. Motivation

People are spending more time online be it for shopping or entertainment and a way to help people save time from searching and selecting aroused our interest. We have all been in situations where apps or websites recommend content you are not interested in or where most of the content you're searching for is not what you want. As a result, we decided to analyze this classification problem for our project and compare different Recommender System and find which approach predicts more accurately.

We will use the Content based, Collaborative filtering and the hybrid Recommender System, which combines two systems together. These systems will provide recommendations for a type of content based on user or content attributes such as rating, genre and tags. This can help users avoid wasting time on checking if the content is relevant to their preferences and also improve their experience. This problem is important because it can provide a better experience to users while at the same time benefiting companies with more consistent users. We have chosen to use Movielens dataset to train, test and improve our methods.

III. Method

Our approach for predicting a likeable recommendation will be done by discerning similarities, within our dataset, between either users through their movie ratings feature or between movies through their tags and genre features. This allows us to proceed with either a content based recommender, a collaborative filtering based recommender or a hybrid of the two.

Since the Movielens dataset is composed of primarily categorical based features we need to assess them in a way such that the training algorithms will be able to use them accurately. Therefore we can use categorical binning, one-hot encoding or multi-classification methods to improve the training data.

For identifying how similar users or movies are, calculating the cosine similarity between features is a simple and effective approach we are considering as it can be used for both content based and collaborative filtering implementations by using a KNN algorithm.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

For more complex analysis we plan to create an interaction matrix from all the movielens training dataset to allow for more detailed algorithms to process the data through approaches such as matrix decomposition and SGD for training and testing. Interaction matrices also allow us to better understand the quality of the data as we will be able to measure its completeness and potentially fill it in as our implementation learns from the training sets.

We can then also introduce some similarity weighted features to these approaches by weighting the datasets calculated genome scores and genome tags that will allow us to more accurately find the most relevant similar entries and predict how likely a movie recommendation will be for a given user.

Finally, we can then compare the test results of these algorithms to accurately judge each implementation's recommendations through an accuracy score derived from these datamodels results.

IV. Intended Experiments

Our experiments will focus on the implementation of KNN and SGD. The dataset will first be normalized and set up with features based on the content and split 80/20 for training and testing. It will then be put into an interaction matrix, which is

created to allow more detailed algorithms to process data and also for measuring the quality of the dataset.

We will then train and test out the content based and collaborative filtering approaches using KNN and SGD, and then compare the previous two methods with a hybrid approach that will leverage the strengths of both. Once implemented and trained, we will test these methods by comparing their results in terms of recommendation accuracy.

Then we will validate the accuracy of the content based recommendations through a scoring system. The scores will be determined by comparing the result of the training dataset with the testing dataset to measure the accuracy of the validation. At last, we will conduct a series of tests with a different training dataset to determine the overall performance of the algorithms against each other.

V. Planning and Milestones

A. Planning

The project will be done in a collaborative fashion, each member performing sections of each step together so as to ensure equal knowledge amongst all members. As this course is being done remotely, this approach done online through weekly meetings will facilitate out plans easier.

B. milestones

1) Milestone 1 (Mid October): MovieLense data set clean up and feature categorization done by all members

2) Milestone 2 (End October): Create a streamlined training and evaluation pipelines for the content based, collaborative filtering and hybrid system approaches done by all members

3) Milestone 3 (Mid November): training and testing each recommender approach done by all members

4) Milestone 4 (First Week of December): Result analysis and scoring between the created system approaches done by all members

5) Milestone 5 (10 December): Writing the report and preparing the 5-minute video done by all members

References

[1] G. Seif, "An easy introduction to machine learning Recommender Systems," KDnuggets, Sep-2019. [Online]. Available: https://www.kdnuggets.com/2019/09/machine-learning-recommender-systems.html. [Accessed: 04-Oct-2021].

[2] P. Kordík, "Machine learning for Recommender Systems - Part 1 (algorithms, evaluation and cold start)," Medium, 15-Dec-2019. [Online]. Available: https://medium.com/recombee-blog/machine-learning-for-recommender-systems-part-1-algorithms-evaluation-and-cold-start-6f696683d0ed. [Accessed: 04-Oct-2021].

[3] "Movielens," GroupLens, 02-Mar-2021. [Online]. Available: https://grouplens.org/datasets/movielens/. [Accessed: 04-Oct-2021].

[4] M. Mahowald, "Building a recommender system," KDnuggets, Apr-2019. [Online]. Available: https://www.kdnuggets.com/2019/04/building-recommender-system.html. [Accessed: 04-Oct-2021].