Recommender Systems Summative

Thomas Pettit   
Computer Science  
Durham UniversityDurham  
thomas.w.pettit@durham.ac.uk

*Abstract*—This paper explains how recommender systems can be used in the domain of film recommendation to create personalized recommendations to active users, using both content-based and collaborative filtering approaches. These two approaches are then evaluated using MRSE and a measure to determine serendipity of the recommendations. The dataset used is the MovieLens 25 Million dataset [1].

Keywords—Recommender Systems, MovieLens Dataset, Content-Based Filtering, Collaborative Filtering

# Introduction

## Domain of the Application

The domain of this application is movie/film recommendation. In particular, providing personalised recommendations to an active user of the system.

## Purpose and Aim of the Application

Movie and film streaming is a market that is growing very quickly, with an estimated market share of 60 billion dollars worldwide, and is predicted to reach 220 billion dollars by 2028 [2]. Therefore, it is in streaming companies’ interest to improve their service in order to maximise their portion of this market share. One way in which companies can get ahead of their competitors is to provide better and more personalized recommendations to their users, as this means their users will be more inclined to stay on their platform to have access to these improved suggestions, instead of heading to a competitor platform. With the huge amount of data that companies such as Netflix and Amazon Prime Video now have on user ratings, and genres and tags for each film on their platform, their ability to provide such recommendations has increased vastly, and so recommender systems that harness this data are improving quickly.

The aim of this application is therefore to demonstrate how, using large amounts of data, recommendations can be provided to the active user, using content-based and collaborative-filtering approaches, that maximise both accuracy and serendipity.

# Methods

## Data Description

The dataset used is the MovieLens 25 Million Dataset [1]. This dataset includes 5-star rating and free-text tagging activity from MovieLens, a movie recommendation service.

The dataset includes 25000095 ratings, and 1093360 tags across 62423 movies, from 162541 different users. In order to ensure that there is enough ratings data for training the recommender systems for each user, each user in the dataset has rated at least 20 different movies. In order to hide the identity of the users, only an ID can identify each user, which has no reference to their actual identity elsewhere.

The dataset itself contains various files, namely: genome-scores, genome-tags, links, movies, ratings and tags. In this paper, the genome-scores, genome-tags, movies, ratings and tags files were used.

The dataset includes a Tag Genome. This is a data structure that contains the relevance of each tag to each movie, with each movie having a value for every tag in the genome. The tags indicate a range of features for a film, including: atmospheric, thought-provoking, realistic and more. The genome itself was calculated using a machine learning algorithm, trained upon user-contributed content such as tags, ratings and textual reviews, as described in [3]. This Tag Genome is represented by the files: genome-scores.csv and genome-tags.csv. The genome-scores.csv file contains the calculated relevance of each tag to each movie, and the genome-tags.csv file provides the description of each tag from each tag ID.

The movies file contains the title and genres associated with each movie ID.

The ratings file contains all the ratings given by users on the movies, as well as a timestamp indicating when the rating was left on the given film, represented in seconds since midnight Coordinated Universal Time (UTC) of January 1, 1970.

The tags file contains all the tags given by users on films, as well as the timestamp that the user gave the tag for the corresponding film, also formatted in seconds since midnight Coordinated Universal Time (UTC) of January 1, 1970.

## Data Preparation and Feature Selection

The datasets themselves contained no missing values in their columns. No pre-processing had to be done on any of the data files.

There were 1129 tags for each movie, each with their own relevance score for each movie. All of these tags were used to train both the content-based and collaborative-filtering recommender systems. This is because the more tags, the more accurate the recommender systems can be, as they can learn from a greater volume of data to learn user preferences from. The time to making predictions for the content-based recommender is also reasonable, and the time to train the model for the collaborative-filtering model is also not too long for this many tags.

## Recommendation Algorithms

Content-Based Filtering Approach

Collaborative-Filtering Approach

In this approach, a model-based collaborative-filtering recommender system is created, based upon a Matrix-Factorisation algorithm, specifically SVD, popularized by Simon Funk in the Netflix Prize. This approach works in the following way:

The rating for an item i by user u, is given by the following formula:

where 𝜇

## Evaluation Methods

# implementation

## Input Interface

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections A-D below for more information on proofreading, spelling and grammar.

## Output Interface

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

# Evaluation

## Evaluation Metrics’ Results

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

## RS Techniques’ Comparison

# conclusion

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4: 19:1–19:19. <https://doi.org/10.1145/2827872>
2. Research ltd and Markets. Global video streaming market size, share amp; trends analysis.
3. Jesse Vig, Shilad Sen, and John Riedl. 2012. The Tag Genome: Encoding Community Knowledge to Support Novel Interaction. ACM Trans. Interact. Intell. Syst. 2, 3: 13:1–13:44. <https://doi.org/10.1145/2362394.2362395>
4. K. Elissa, “Title of paper if known,” unpublished.
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.