Article Recommendation System

Himali Goel(160020002), Manika Khare(160020020) and Siddharth Singh(160010027)
Team 07

November 25, 2018

1 Introduction and Overview

Recommendation systems represent user preferences for the purpose of suggesting items to purchase or examine. They have become fundamental applications in electronic commerce and information access, providing suggestions that effectively prune large information spaces so that users are directed towards those items that best meet their needs and preferences.

A variety of techniques have been proposed for performing the task of recommendation, including content-based, collaborative, knowledge-based and other techniques. While collaborative filtering makes automatic predictions (filtering) about the interests of a user by collecting information regarding preferences or interests of other users (collaborating), content based filtering uses information about the description and attributes of the items that the user has previously consumed to model user's preferences. However, Recent research has demonstrated that a hybrid of the two models could be more effective than the pure approaches since it combines the benefits of the two, overcoming some of their shortcomings. Our project aims at building one such hybrid recommendation system to suggest news articles to readers based on their prior interactions with previous reads in the form of views, likes, comments, bookmarks and followings.

Related work

Recommender systems are systems with the ability of providing suggestions or directing a person to a service, product or content, that has a potential of interest among a number of different alternatives. Currently, highly rated internet sites as Amazon, eBay or YouTube use recommender systems as part of their services. Beyond these examples, recommender systems also apply on numerous other domains including books, movies and TV programs, music, news articles and so forth. Over the years, recommendation of online news articles has become an area of great interest. For instance, large newsfeed portals, such as Google News, and Yahoo! News, provide personalized news recommendation services for a large amount of online users.

The task of recommending news articles based on the user's preferences, can be conducted using distinct methodologies. Approaches adopting content-based and collaborative filtering are widely used by existing news recommender systems. Some content-based news recommender systems have been proposed in the last decade. An example is NewsDude that presents news stories to the user, who then rates the articles according to whether they are interesting or not. The user profile is then compared with content of other news stories to generate personalized recommendations.

Collaborative recommendation systems consider that users with similar reading behaviours in the past usually will have similar preferences about news articles in the future. Google News is a popular example of a news articles recommender based on collaborative filtering. Google News is an online news portal that aggregates news articles from thousands of sources, grouping them to the users, according to their personal interests.

Content-based and collaborative filtering can provide meaningful recommendations. However, each of the approaches have however some disadvantages. In order to improve the performance, hybrid approaches to news recommendations have also been explored. Representative examples include P-Tango which presents a hybrid approach that recommends news items by combining content-based and collaborative filtering recommenders together using a weighted average function.

2 Methods

2.1 Popularity-Based Model

We use the popularity-based model to recommend to users with less than 5 interactions. This model is not personalized since it simply recommends to a user, the most popular items that the user has not previously consumed. As the popularity accounts for the "wisdom of the crowds", it usually provides good recommendations, generally interesting for most people.

2.2 Content-Based Filtering

Content-based filtering approach uses the description or attributes of items that the user has previously interacted with in order to recommend similar items. For textual items, like articles, news and books, we can simply use the raw text to build item profiles and user profiles. We have used a popular information retrieval technique named TF-IDF for performing content-based filtering. This technique converts unstructured text into a vector structure, where each word is represented by a position in the vector, and the value measures how relevant a given word is for an article. In our model, we consider both unigrams and bigrams as features. As all items will be represented in the same Vector Space Model, we can use cosine similarity to effectively calculate similarity between different items.

2.3 Collaborative Filtering

This method makes automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating). It tries to make recommendations to a user based on the preferences of other similar users. In

our project, we have used a model-based technique called Singular Value Decomposition (SVD). Latent factor models compress user-item matrix into a low-dimensional representation in terms of latent factors. One advantage of using this approach is that instead of having a high dimensional matrix containing an abundant number of missing values we will be dealing with a much smaller matrix in lower-dimensional space. comparing similarity on the resulting matrix is much more scalable especially in dealing with large sparse datasets.A crucial element is the number of factors to be used to factor the user-item matrix . The higher the number of factors, the more precise is the factorization in the original matrix reconstructions. Therefore, if the model is allowed to memorize too much details of the original matrix, it may not generalize well for data it was not trained on. Reducing the number of factors increases the model generalization. After the factorization, we try to to reconstruct the original matrix by multiplying its factors. The resulting matrix is not sparse any more. The matrix generated represents the predictions for items the user have not yet interacted with, which we will exploit for providing recommendations to the user.

2.4 Hybrid System

In fact, hybrid methods have performed better than individual approaches in many studies and have being extensively used by researchers and practioners. Let's build a simple hybridization method, by only multiply the CF score with the Content-Based score, and ranking by resulting score.

Datasets

We will be using the Deskdrop dataset, which contains a real sample of 12 months logs (Mar. 2016 - Feb. 2017) from CIT's Internal Communication platform (DeskDrop). It is composed of two CSV files:

- shared_articles.csv: Contains information about the articles shared in the platform.
- users_interactions.csv : Contains logs of user interactions on shared articles.

Results

In Recommender Systems, there are a set metrics commonly used for evaluation. We chose to work with Top-N accuracy metrics, which evaluates the accuracy of the top recommendations provided to a user, comparing to the items the user has actually interacted in test set.

For each item, a particular user has interacted with, we sample 100 other items that the user has never interacted with thus creating a set of 101 items. A rank list of recommended items using the model to be evaluated is then constructed from this set.

The Top-N accuracy metric chosen was Recall@N which evaluates whether the interacted item is among the top N items in the ranked list of 101 recommendations for a user. The following results have been recorded for different models:

- Popularity-Based System:
  - Recall@10 : 0.37
  - Recall@5 : 0.24
- Content-Based Filtering :
  - Recall@10 : 0.53
  - Recall@5 : 0.42
- Collaborative Filtering :
  - Recall@10 : 0.45
  - Recall@5 : 0.33
- Hybrid System :
  - Recall@10 : 0.55
  - Recall@5 : 0.45

We observe that the hybrid system has the highest accuracy among all models, which leads us to conclude that it is better for recommendations in this case in comparison to the individual approaches of content based and collaborative filtering.

3 Discussion and Future Directions

The interest on online newspapers has been growing significantly over the past years. In order to present the most relevant news articles to users, different recommendation systems have been made available using various techniques in order to make access to large amounts of information more efficient.

Future work includes evaluating the performance of our implementation using a larger set of users and articles in a real environment. Another experience may involve changing or switching the priority value and the percentage that each of the recommendation approaches (content-based and collaborative filtering) provide to the final list and analysing the impact. We can further add an aspect of time and location which updates the user preferences based on the recent interests and locality of the user. With these analyses, the model will further be able to detect how the user reading behaviour changes and try to infer user preferences with more accuracy .

References

[1]https://www.researchgate.net/publication/303772274_A_hybrid_recommendation_system_for_news_in_a_mobile_environment#pf8

[2]https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[3] https://docs.scipy.org/doc/scipy/reference/generated/scipy.sparse.linalg.svds.html