Twitter-based Recommendation System

Jiana Feng jf3283
Ziying Liu zl2839
02/14/2020

Abstract User behaviors on social media imply tremendous potential information. By analyzing user behaviors, we can deduce a large amount of potential interests from users and support users to make decisions. In this project, we will build a Recommendation System that can provide Twitter users with personalized tweets home pages, trends and similarly interesting accounts. We utilize content-based, collaborative filtering techniques, TF-IDF and k-means clustering to generate a list of recommendations.

Keywords Twitter; Collaborative Filtering; K-means Clustering; TF-IDF

1 Background

Twitter is formed by a recommender system because most of the twitter users follow more than 80 people, which means if the system just shows the latest tweets at the top, some important news or high-quality tweets will be overwhelming by trivial information. To solve the information overload problem, every news and tweet you read is selected by the recommendation algorithm, which is an implicit system. 'Trends for you', 'Who to follow' and advertisement are traditional recommender systems.

Collaborative filtering is the most popular algorithm in the recommender systems. This approach is based on the collection and analysis users' behaviors, activities and preference. A key advantage of this approach is that it does not require creation of explicit user profiles and specific domain knowledge.

2 Introduction

The goal of our project is to build a complete recommender system that can detect the users' potential interests . After a user log in, which means uploading his history tweet into our system, we can provide them with personalized tweets home pages, some popular news with hashtags and the people that they might be interested in.

The first step is collecting enough datasets from a connected network, which is because we want to finally generate their home page. Based on users' history behavior, such as following another user, publishing a tweet, and retweeting, to build the user profiles, or in another word, interest tags. We will use TF-IDF and some other text analysis to extract those tags, connect them to each unique user ID and store them in the system cache. Those tags will help us to find the targeted group for certain advertisements as well.

The traditional recommender system will treat a tweet as an item, and retweeting or liking as criteria to construct the User-Item matrix. However, it will cause extreme sparsity of the UI matrix, which is the main obstacle for the recommendation. In our project, we will construct a User-Term(Tag) matrix and User-Followee (people you are following) matrix to calculate the similarity and make a recommendation for the user home page. The most efficient way to evaluate this approach is to calculate the average precision in our validation set.

As for hashtag recommendation, we plan to use the Hashtag Frequency- Inverse Hashtag Ubiquity(HF-IHU), which is a variation of the TF-IDF. It can consider hashtag relevancy, as well as data sparseness. HF-IHU can find the most popular hashtags and we will use the user profile to analyze the similarity with the top 50 hashtags. Select the most suitable popular hashtags for the user. For this method, we examine the precision and recall in the validation set. However, because we want to recommend new hashtags to our users, the precision might be kept low.

We build a user-based collaborative filtering algorithm which determines highly personalized suggestions. We will compute the similarity between users and provide suggestions of accounts or tweets to users based on the contents that other highly similar users have focused. We plan to use Pearson Correlation method to estimate the similarity between users.

3 Dataset

We will extract Twitter review dataset via Twitter API. We randomly select one user and extend other users through this user’ s followers and fans. By using python, we will collect around 10,000 users and download their tweets and individual information. Since the instance of Twitter recommendation, tweets that we collect will be split as train dataset, validation dataset and test dataset. After finishing our system, we will upload the latest tweet within 30 minutes and recommend each user in test set a personalized homepage, hashtags and interested account recommendations.

4 Plan

- milestone 1: Background. Project goals. Dataset.
- milestone 2: Data scraping by python. Data cleaning. Text analysis by using TF-IDF. Building up the users' profiles.
- milestone 3: Construction of matrix of a User-Term(Tag) and User-Followee. Generation the top 50 hashtags based on similarity.
- Final: Model selection. Decision on Final system. Evaluation of outputs.

Reference

[1] Pengfei Z & Zhijun Z. (2019). Collaborative Filtering Recommendation Algorithm Integrating Time Windows and Rating Predictions, Applied Intelligence, 49, 3146-3157

[2] Nitin P.K & Zhenzhen F. (2015). Hybrid User-Item Based Collaborative Filtering. Procedia Computer Science, 60, 1453-1461.

[3] Dheeraj B, Sheetal G. (2015). Matrix Factorization Model in Collaborative Filtering Algorithms: A Survey. Procedia Computer Science, 49, 136-146.

[4] F.O Isinkaye & Y.O Folajimi. (2015). Recommendation System: Principles, Methods and Evaluation. Egyptian Informatics Journal. 16, 261-273

[5] V. Rakesh, D. Singh, B. Vinzamuri, and C. K. Reddy (2014). Personalized recommendation of twitter lists using content and network information. In ICWSM.