Social Relationship Recommender

CMPE 256 Large Scale Analytics

Apoorva Shadaksharappa

Harshitha Yerraguntla

Mrinalini Gopalakrishnan

Smitha Eshwarahalli Ramesh

Venkatesh Tanniru

Abstract:

Social platforms play a vital role in online communication, content sharing, and other business applications. They collect huge amounts of data regularly from their users. Data collected from users is reused to recommend their applications to the users. Recommender systems have several applications such as product recommendation, human recommendation, place recommendation, job recommendation, etc. Social media uses various methods for recommending friends to a user, Facebook - uses collaborative filtering which makes use of the data set made of more than a billion users and 100 billion user ratings. Instagram - recommends friends based on attributes such as keywords, hashtags, like, comment, share which reflect the user interest in certain topics and posts. Twitter - recommends users based on the accounts they follow, retweets and favorite tweets. The main objective of this paper is to recommend followers on Twitter by analyzing hashtags and tags in the user tweets data. The dataset is taken from Twitter developer API with groups of accounts and users, tweets and replies using developer tokens of Twitter. The system is developed in two phases, i.e. In phase 1 we are loading the user and their attributes into a data frame for the calculation similarity matrix using Pearson's correlation and in the next step we finding the nearest neighbors for each user and recommending the users whom to follow. In phase-2. Once we recommend the users whom to follow, we evaluate our recommendation based on the concept of Reinforcement Learning, where the not so important features are extracted and assigned less weightage.

Introduction:

Nowadays people tend to communicate or share information in the virtual world with the help of the internet. These virtual worlds are known as social media platforms. Social media platforms collect a huge amount of data from their users and use the same data to connect users with people/businesses via recommendations. These recommendations help them to build their business. Users previous engagements, interests, familiarity, content shared, platform usage and several other attributes are considered for generating personalized recommendations. Recommendations help to make efficient use of the application and help to build a better user interface of the application. Recommender systems have several applications such as product recommendation, human recommendation, place recommendation, job recommendation, etc. This paper concentrates on human recommendation. Human recommendations help to expand and strengthen human relationships. Countless studies have demonstrated that we are happier when we have strong social relationships. Social Platforms need a strong recommendation algorithm that builds meaningful relationships between its users. Several old generation recommendation systems just recommend using the similarity score calculated between users.

There is a need for a new efficient recommendation system.

The main objective of this paper is to develop a recommender system that recommends followers on Twitter with better efficiency than regular recommendation systems. The Proposed recommendation system recommends using a collaborative filtering approach and dealing with Twitter data offline. The goal of the proposed system is to identify the most meaningful relationships and recommend them.

Jack Dorsey developed Twitter in October 2006. Twitter is one of the major social media platforms for sharing information. It allows 140 characters per message known as a tweet. Tag is called as the user id given to every user with the character'@' preceding it. If one user posts a message another can share the message through a retweet. Trending topic with a word and a character "#' in front is called a hashtag. Relevant topics are sorted using these hashtags. Twitter uses these tweets, retweets and favorite tweets for generating personalized recommendations.

Related Work:

Techniques used for building recommendation systems are Content Based Filtering (CBF), Knowledge Based Filtering (KBF), Collaborative Filtering (CF) and Hybrid Filtering (HF) [1].

Content based filtering algorithm approach uses the content that the user accessed in the past. This algorithm also considers user profiles to consider what the use user likes and recommend similar items what user likes [2]. CBF considers some set of attributes to describe each item and classify according to the preferences. This algorithm is mostly used for recommending text documents.

Knowledge based filtering algorithm approach uses the inferences of user's needs and preferences for recommending the items. It uses functional knowledge of how the item is useful for the user. This algorithm is mostly used for recommending movies and clothes.

The hybrid filtering algorithm approach combines collaborative filtering and content-based filtering. Hybrid Filtering could make use of both the content as well as the similarities among users for generating recommendations. This algorithm is used to overcome the disadvantages of both content-based and collaborative filtering. This algorithm makes predictions based on a weighted average of the collaborative recommendation and the content-based recommendation. Netflix uses a hybrid filtering algorithm for generating recommendations.

Collaborative filtering

Collaborative filtering is the most commonly used technology for human recommendation. This Algorithm is based on users' past preferences and find similarity between the users. Collaborative filtering works on by building a database (user-item matrix) preferences for items by the user. Then it matches users with relevant interests and preferences by calculating similarities between their profiles to make recommendations. CF approaches assume that those who agreed in the past tend to agree again in the future [3].

Recommendation systems with collaborative filtering models which assume that people like things that are similar to the things they like, and things that are liked by people with similar taste. Collaborative Filtering is categorized into two types, i.e. memory based Collaborative Filtering and model based Collaborative Filtering.

A. Model based Collaborative Filtering

Model based on collaborative filtering consists of models that are developed using machine learning algorithms to predict the user's rating of unrated items. This algorithm is based on an offline pre-processing or "model-learning" phase. At run-time, only the learned model is used to make predictions. The models are updated and re-trained periodically. Building the model and updating it regularly can be computationally expensive. This algorithm uses a large variety of techniques. Machine learning algorithms used are Clustering based, Matrix factorization and Deep learning. This type of algorithm is used for huge sparse data. This type of filtering algorithm has advantages such as Scalability, fast prediction speed, and low overfitting. These types of algorithms are inflexible and less quality of prediction.

B. Memory based Collaborative Filtering

Memory based algorithms, also known as the neighbor based algorithm. Memory based algorithms are widely used in many of the social media platforms and e-commerce sites, such as Amazon, Facebook, Twitter, etc. The Memory-based algorithm approaches is to find a correlation between the users and items in the database to predict the similarity score between them. This predicted score is used whether the user likes the item in the future or not. Using this score, we can recommend the item to a user or not. This algorithm uses an entire database to predict a similar product that the target user might have liked or rated or used. This method consists of two types of algorithms, i.e. item based, and user-based filtering.

- ***Item Based Collaborative filtering***

Item based collaborative filtering methods calculate the similarities within a group of items that are rated or searched by the user. item based methods are used for recommending similar products based on the products already used or searched by the user. This model specifies the target user and finds the items rated by the user. It also predicts similar items to the user rated items and recommends them.

- ***based Collaborative filtering***

User based methods predict the recommendations based on the similarity between the user. In user-based, users who gave ratings for items in the past will tend to give a similar rating in the future. The algorithm considers the item rating of the user and then predicts the user's rating for the item for which the user has never interacted. Users preferences remain stable and throughout execution. This model specifies the target user and finds the items rated by the user. Then extracts the items that the user has never used and recommends them to the user.

- ***Graph based recommendation.***

Although there are no graph-based methods designed for collaborative ranking, many recent studies have been conducted in other areas of recommender systems. Here, we will briefly review those algorithms and clarify the main differences between the current work and them.Graph-based recommendation algorithms are composed of two steps: Constructing a graph representing the data and making recommendations by analyzing the graph. These recommendation algorithms have exploited different types of graphs. However, in all of them, the main component of the graph is the relations between users and those items that have been rated by them. Therefore, the most common approach is constructing a bipartite network where the connections are from one part of the network, users, to the other part, items. Once the bi-partite graph is constructed, several approaches can be used to rank the items using the information from the neighbors of the target user. Approaches like using common neighbors, Katz similarity, diffusion scores and personalized PageRank have been used in this domain

Recent methods have extended the bi-partite network by adding some layers to it. Some researchers (Xiang et al., 2010) have considered using a session layer to take into account the long-term and the short-term preferences of the user in order to make recommendations in a particular time.

Implementation

A. Data Extraction

Unlike other Social Platforms, almost every user's data is public on Twitter. The dataset is obtained from Twitter developer API with groups of accounts and users, tweets and replies using developer tokens provided by Twitter. Using Twitter developer API, the user data is scrapped daily into the cloud. Trending hashtags and tags are considered as binary attributes for scraping the data. The data is scrapped based on the geographical location and trending topics. The scrapped user data is stored in a CSV file in JSON format.

B. Preprocessing Data

From the dataset extracted from the Twitter developer API, Hashtags and tags are fetched. This fetching includes cleaning and tokenizing the tweets for or Fetching trending topics/features (hashtags) and user list. Redundant attributes such as #happy and #happiness, which give the same meaning are eliminated and only one is considered. The relationship between attributes have similar meaning most of the time E.g. #love and #happily_appear_together, we use just one attribute instead of two.

Reinforcement learning is an area of Machine Learning. Reinforcement. It is about taking suitable action to maximize reward in a situation. It is employed by various software and machines to find the best possible behavior or path it should take in a specific situation. Reinforcement learning differs from the supervised learning in a way that in supervised learning the training data has the answer key with it so the model is trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task. In the absence of training dataset, it is bound to learn from its experience.

Special Cases:

The collaborative filtering (CF) approach in recommender systems assumes that there is enough and consistent information is available for each user. In some special cases and for some users this approach fails. There are two types of special cases detected in human recommendation, they are Cold start and Gray sheep. Fig.1 is used for explaining these special cases

- **Gray sheep**

Some of the twitter users are unique that they don't interact or follow others and they don't do regular tweets or retweets.

From Fig.1, U6 and U7 are unique with others and they just have a connection with each other. These types of users have a unique taste and they don't interact with others. In these types of special cases, we use a technique called gray sheep to recommend such unique followers to the user. To solve this problem we random people and store their feedback. This feedback undergoes reinforcement learning and generate personalized results.

- **Cold Start**

When generating recommendations to a new user, it is difficult to generate recommendations as he doesn't have any previous data and the system doesn't know any of his preferences. From Fig.1, U8 doesn't have any friends or data available. When a user newly creates a Twitter account, he doesn't have any specific preferences. We need to recommend some users assuming his/her preferences and similarities. We can just recommend some of the most popular Twitter celebrities to him.

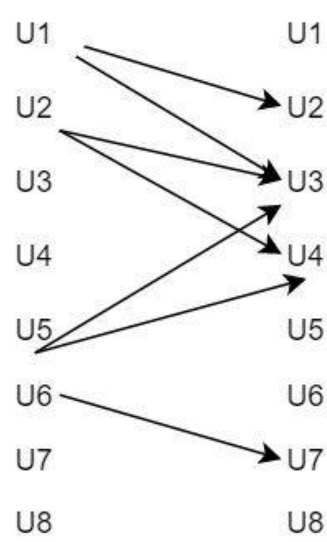
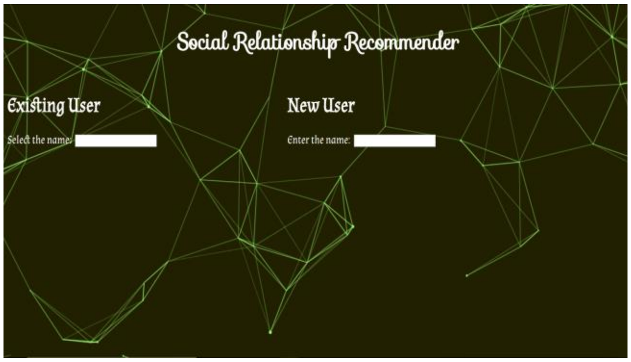
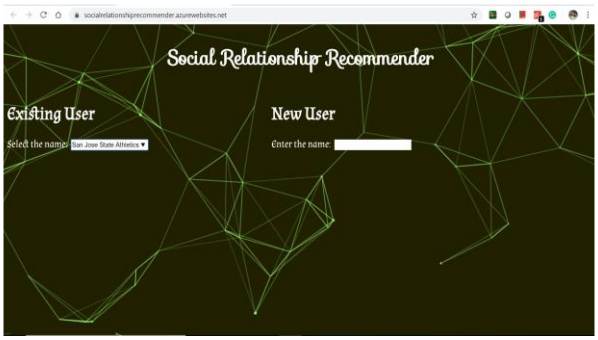
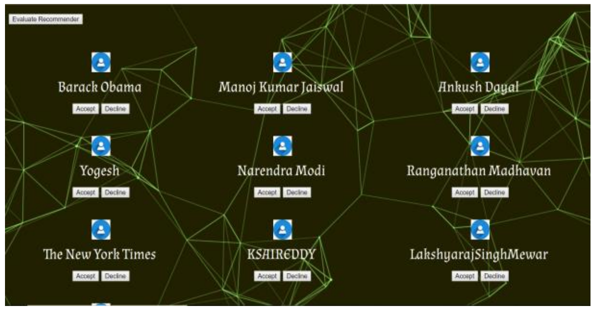


Fig-1. Special Cases for recommendation

Results:







Conclusion:

Twitter follower recommendation system was implemented with the scrapped tweet data. The KNN model output was verified for the twitter data. The output from the KNN model is used for recommending followers using collaborative filtering. The recommendations generated are evaluated using MSE and MAP metrics for accuracy. The collaborative filtering (CF) approach in recommender systems assumes that users' preferences are consistent among users. Although accurate, this approach fails on some users. We presume that some of these users belong to a small community of users who have unusual preferences, such users are not compliant with the CF underlying assumption. They are grey sheep users. Cold start problem in recommender systems can be relevant both for new users of some service and for a new product, which has yet no reviews or history of being a success among a certain group of users. And in case there aren't enough user actions for a item, the engine will not know when to display it. This problem is countered by suggesting the user with famous personalities. Further research needs to be done on developing an efficient recommendation system with more attributes to improve accuracy.

References

[1] Najdt Mustafa ; Ashraf Osman Ibrahim ; Ali Ahmed ; Afnizanfaizal Abdullah, "Collaborative filtering: Techniques and applications," *International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE)*, 2017.

[2] Prateek Parhi, Ashish Pal, Manuj Aggarwal , "A Survey of Methods of Collaborative Filtering," *International Conference on Inventive Systems and Control*, 2017.

[3] Xiangwei Mu ; Yan Chen ; Taoying Li, "User-based collaborative filtering based on improved similarity algorithm," *International Conference on Computer Science and Information Technology*, vol. 3, 2010.