CS-5312 Big Data Analytics

SRUTTA - Stream Rating Using Temporal and Textual Analysis

Ahmed Muqaddas {16030039}, Talha Khan {18100267},
Syed Mohsin Bukhari {16030038}

1 Problem Description

Recommendation systems have become a necessary feature for all the websites on Internet due to their commercial demand. These systems are useful for both, businesses and their clients. They provide a platform where sellers can directly reach people who are most likely to buy their products. On the other hand, buyers are presented with suggestions to purchase items based on their own profile. This can help the buyers avoid unwanted items and quickly discover what they might want.

These recommendation systems usually only consider the number of times a buyer has interacted with similar items before recommending an item. This can further be extended to include recommendations regarding items that other similar buyers have purchased. Another way of generating recommendations is by analyzing the sentiment contained in the other buyer's reviews regarding items.

The method used for recommending items by combining buyer reviews with the maturity of buyers is relatively unexplored. Also, traditionally these items and the reviews regarding these items change very slowly over time. However, if these items are streams on social media then the sentiment is likely to change very rapidly based on the content of stream. The problem of recommending these streams to users is also relatively new idea, yet an interesting one because these streams have only started to appear on social media recently as compared to other online content. The idea is to assign ratings based on the posts by other similar users and the current users history.

INPUT		OUTPUT
list of streams	→	sorted list of streams with ratings

2 Motivation

The amount of information and content broadcasted over social media in form of streams is rapidly increasing. However, the lack of a proper framework which can recommend these streams to users poses a challenge to the research community to find a way to fill this gap. Users usually have to sift through multiple pre-determined categories of these streams to look for the ones that might interest them.

On the other hand companies like ***Twitch*** and ***Twitter*** want users to keep coming back to their websites by providing them with the kind of entertainment or content they prefer. This will simultaneously give power to the users for exploring the streams with ratings that are crafted specifically for the user.

Aim of this project is to develop a framework that provides users of streaming websites, with a platform where they can see all the streams along with the predicted ratings for these streams. Users might want to sort the streams in opposite order of sentiment so they can choose to have a look at the content that goes against their views. This can help them identify streams that they can target in order to spread their own views through chat or reviews.

3 (Potential) Tools and Techniques

The power of prototyping that ***Python3*** provides along with its widely used libraries makes it the first choice for this project. The ***NLTK*** (Natural Language Processing Toolkit) used in conjunction with others like ***gensim*** provides a great number of options for analyzing user generated textual content on online social media websites.

In case, a user logs in to a streaming website, ***Twitch***, a user might want to join a stream that is at the highest point of entertainment. It can be really useful to incorporate this information when predicting ratings for streams. The highest point of entertainment might be correlated with the time for which the stream has been online. It can also be useful to incorporate user related temporal statistics, like the time for which the user has been currently active. If the time matches with the time for which stream has been online then the user might be more likely to watch that stream.

A user, on ***Twitch*** for example, might be very similar to another user based on their chat messages on the streams. Extraction of the sentiment hidden in the chat messages by users can be of real importance when determining a rating for the stream. This can be achieved by learning feature representations of chat messages and then performing regression to obtain a rating.

The existing techniques for recommendation systems do not take into account the combination of temporal and textual properties of related content for performing collaborative filtering. This project will focus on combining and modifying the methods introduced previously to take these data features into account. The time at which a stream was started the time at which a user logged in, might be a very strong indicator of whether the user will join the stream or not. This can help produce a rating for the stream and it can be combined with another sentiment analysis based rating to produce a final rating.

The consideration of features mentioned above has potential for improving recommendation systems for streaming data. The time for which a user is active along with the time for which the stream has been broadcasted are unique features in the class of data that is constantly pouring in, for example in streams.

4 Related/Background Work

There is huge amount of literature available for recommendation systems. However, the background work related to the kind of data features that we are using is very limited when the data is coming in continuously.

In order to explore the usefulness of methods for comparing temporal features, the publication `BiCycle: Item Recommendation with Life Cycles` is great point to start. However, for the use of textual features to predict sentiment rating from user reviews, the resources available on Internet are quite extensive. A particularly useful resource regarding Nature Language Processing is the playlist of video lectures of `Deep Learning with Natural Language Processsing` on YouTube.

Previous works in this field do not combine the features that are needed for data that is continuously being captured. The techniques for merging the ratings obtained from temporal and textual analysis are not used in this context before, therefore, the existing methods might under perform on data that has rapidly changing user sentiment.

5 Your idea and approach

The idea for this project is to build a framework that can provide users with ratings for streaming websites and/or APIs like ***Twitch*** and ***Twitter***. Users will get constantly updated ratings for the streams. In short, when a user sorts streams on ratings, the list will be updated after every few seconds to reflect the rating of streams with respect to the user on runtime.

This will be a very personalized framework for the user which will not surround the user only with things that the framework thinks that user will like. An option to sort and see content that the user will have a negative sentiment about will also be there. This way, the user can customize what they want to see based on the ratings predicted for the user.

This project is limited in a way that it will not be exploring a broad range of features for analysis and generation of user ratings. In fact, only a small subset of possible temporal features will be used for analysis. Also, ratings do not help the content creators in any way. For example, someone broadcasting can not benefit from the outcome of this project.

In future, the techniques used in this project can be used to generate suggestions for content creators to include in their streams so they can attract maximum number of users. Secondly, a parameter optimization technique can be used to find out parameters to combine ratings from different kind of features.

6 Data Sets

Currently, a dataset for ***Twitch*** is going to be used. The shear size of dataset is enough to perform a detailed analysis on it. The compressed size on disk is 12GB, when extracted it takes about 200GB. It is all text data. The data is in *JSON* format and it contains the fields like user chats on various streams along with timestamps. This seems to be enough data for starting work on the project.

7 Evaluation Criteria/Metrics

The design of an evaluation criterion for this project is an intriguing idea. This is because the separation between training and testing data can be done temporally. At any one point, we can predict the ratings for streams that are in future with respect to that point in time. Therefore, everything before this point can be used as training data and everything after this point can be used as validation and testing data.

We can simply use *accuracy*, *precision*, *recall* and/or *F1 measure* to predict if the proposed framework is working as expected. These measures can then be compared with the accuracy measures reported by other techniques.

8 Potential Conference/Publication

The conference that best suits this project is ***ICDM*** (International Conference on Data Mining). The deadline for paper submission is 5th June, 2018. This work is related to this conference because other work published here has great variety in content and the common thing is data. The work in this project is going to include acquiring insights into relatively new kind of data. Therefore, this is a relevant conference. Besides, the techniques described above provide insight into data and this is what the conference is about.