Artist Recommendation System Based on Geographic Location

Ketul Patel
Dept. of Computer Science
University at Albany-SUNY
Albany, USA
kmpatel@albany.edu

Viral Patel
Dept. of Computer Science
University at Albany-SUNY
Albany, USA
vdpatel@albany.edu

Madeeha Khan
Dept.of Computer Science
University at Albany-SUNY
Albany, USA
mkhan8@albany.edu

***Abstract*—Over the past few years, the recommendation systems have played an important role to help users select the preferred product from a massive amount of data. For music recommendation system, most recent recommendation systems made attempts to associate music with the user's preferences primarily based on user ratings to suggest more songs and music according to taste of audience. In this paper, we propose a system that utilizes tweets instead of ratings. Our system recommends the most popular artist of the location. This kind of recommendation proves helpful to event management companies who want to arrange concert of bands, singers to get good business. We have used Twitter API to collect more than 2000 tweets of Albany, New York. We have preprocessed tweets, performed sentiment analysis on the tweets, applied SVM classifier and 10-fold cross validation. After applying association rule mining, we have come up with the top 3 most popular singer or band of the Albany.**

Keywords— Music Recommendation, SVM, Collaborative filtering, band, Sentiment Analysis.

I. INTRODUCTION

With the growth of World Wide Web, a massive amount of music data is available on Internet. A rapid growth of music information enables a large increase in needs of clarifying customers' preferences. Hence, it had been becoming a challenge issue to help customers effectively obtain their preferred music from a huge amount of music data over the past few years. For music recommendation, most recent systems made attempts to associate music with user's preference based on their ratings, However, this kind of recommendation system encounters the problem called rating diversity that makes prediction results unreliable. To cope with this problem, in this paper, we present an approach that utilizes tweets instead of ratings to recommend the artist. Our system is based on Collaborative Filtering Systems, which indicates that we are using historical usage data to recommend items. Moreover, our system recommends the most popular artist or band based on specific location. The applications of this system are:

- Help local artists to understand the taste of the specific area
- Give music suggestion to users
- Music recommendation for clubs, bars, restaurants, etc.
- Music recommendations for local Radio, Television channels.
- Help Event Management companies to grow their business. Event Management companies can know the most popular artist of the specific area and can plan concert of that artist to get more business.

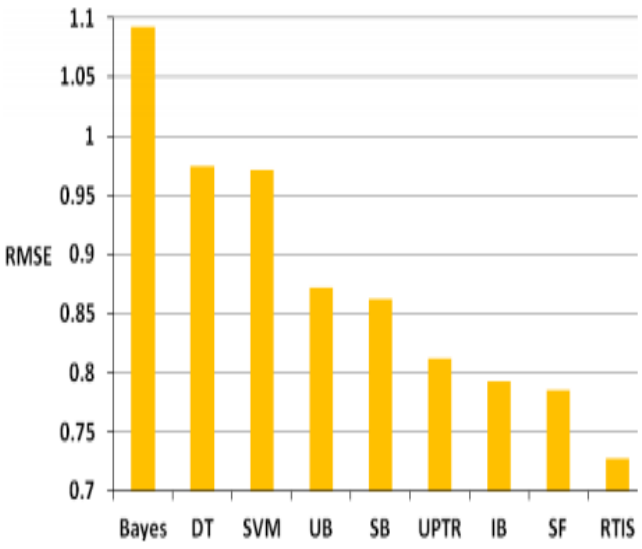
For this system, we have collected more than 2000 tweets related to music of Albany, New York. After collection, we have pre-processed tweets to remove redundant tweets, slang words, unnecessary data, etc. We have performed sentiments analysis on the pre-processed tweets. We have applied SVM classifier and 10-fold cross validation. At last, we have generated our recommendation system using association rule mining.

II. RELATED WORK

A. *Music Recommendation by Mining Social Media Tags[1]*

This paper presents the system that uses social media tags to calculate the similarity between music pieces. The empirical evaluations on real social media datasets reveal that this proposed approach using social tags outperforms the existing ones using only ratings in terms of predicting the user's preferences to music. The authors propose a novel recommender system, namely Recommendation by Tag-driven Item Similarity (RTIS), which adopts play counts as implicit ratings and item tags as semantic preferences. By mining the relations between ratings and tags, the user's preferences can be derived successfully. The experimental results reveal that, their proposed approach can effectively capture the user's preferences on music more than state-of-the-art ones compared with. Moreover, the main aspect behind the experimental results is that, the user's preferences are highly related to social tags. The following figure shows the comparison between RTIS and other recommendation systems.

In real applications, most tag-based recommender systems rely heavily on the tags from social music websites like Last.fm. Unfortunately, social-based music websites merely provide the users' play counts without ratings. It gets the preference prediction type changed from ratings to probabilistic relations. In other words, this type of recommendation mechanisms predicts the relevance between the users and items and thereby generates a ranking list of items (so called Top-N Recommendation, TNR).

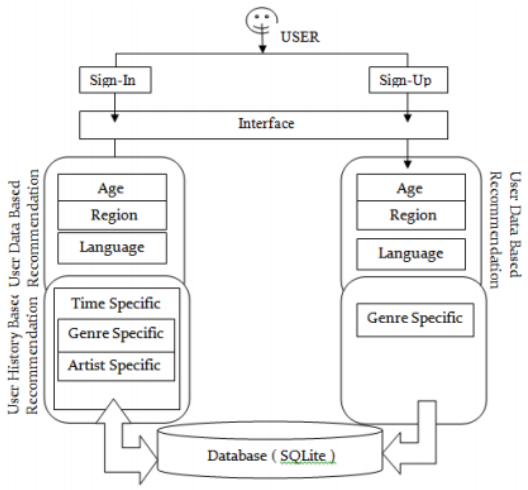


The main contributions of this paper can be summarized as follows: First, item-tag-driven and artist-tag-driven similarities are proposed as good supports in determining the most-relevant items for the prediction. Second, instead of Top N recommendation, they transform play counts into ratings to describe the user's preferences. The experimental evaluations reveal that, this proposed recommender system can bring out the promising results for personalized music recommendation.

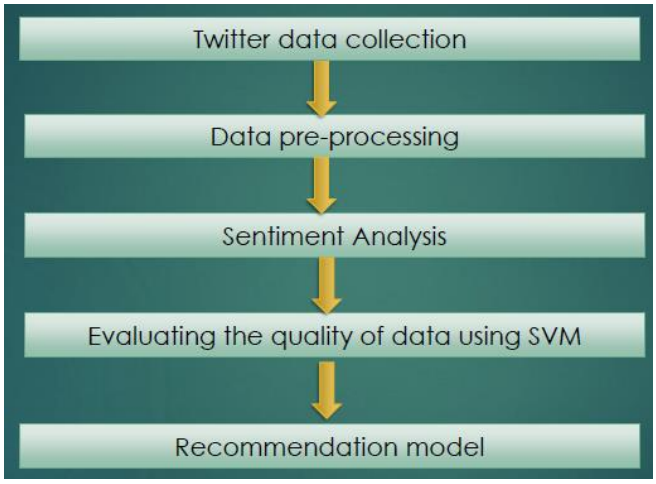
B. *Music-Map by Basundhara Dey[2]*

This system is mainly to help users to filter and discover songs according to their taste, age, region and language. For each music object, the representative track is first determined, and then four features (Genre, Artist, Language and Duration) are extracted from this track. According to the features, the music objects are properly grouped. For users, the personal database table and history are analyzed to derive user interests. The content based, collaborative and statistics-based recommendation methods are proposed, which are based on the favorite degrees of the users to the music groups. A series of experiments are carried out to show that this approach is feasible. The following diagram shows the block diagram of the system.

BLOCK DIAGRAM



III. PROPOSED APPROACHES



The figure above clearly shows the proposed approach for the system. We have collected data, pre-processed data, performed sentiment analysis on pre-processes data, evaluated the quality of the data and generated recommendation model using association rule mining.

We agreed to use Tweets rather than sample dataset for out system because Twitter is the most used social networking application and we can get real time data of the users. To collect the data, we have used Twitter's REST API.

After collecting tweets, the next important step we implemented was to pre-process data. We removed redundant data and other unnecessary data like stop words, links, hashtags, alpha numeric symbols, tweet id, tweet creation time, etc.

After pre-processing tweets, we have implemented sentiment analysis on pre-processed tweets and divided the data into positive and negative tweets.

After performing Sentiment Analysis, we applied SVM classifier on the test set. We have manually labeled half of the tweets as positive or negative and other tweets were labeled by SVM classifier. We have performed 10-fold cross validation to get more accurate results.

We derived association rules for our collected tweets and calculated Support value. Based on the support values, we have built our recommendation system to recommend the most popular artists of Albany, New York.

IV. SYSTEM DESIGN AND IMPLEMENTATION

As we have discussed earlier, recommend an artist or band based on location has many benefits. People always prefer to go to concerts of their favorite artists or bands. So, for event management companies, it is very necessary to understand people's favorite artists, bands, etc. to arrange concert to get more business. Therefore, we have created prototype for achieving good results. We have followed following approach:

- Collecting Tweets
- Data Pre-processing
- Sentiment Analysis
- Applying SVM Classifier
- Association Rule Mining

A. DATASET

We have collected tweets and used tweets as dataset for our system. Twitter is the most popular application amongst users where users share their opinins, personal interest,etc. so we could get real time data of the users. We have ued Twitter's REST API to collect tweets. We have collected more than 2500 tweets. We have collected tweets from Albany location using query related to music.

Twiiter's REST API has limit of 180 tweets per 15 minutes. Therefore, we had to wait for 15 minutes to collect more tweets and run python script every 15 minutes. To solve this issue, we have written python script that will run after 15 minutes so that we do not have to run script to collect tweets every 15 minutes.

The following values we determined:

- API Recall : 6.8
- Quality Precision : 0.74
- Quality Recall : 0.66

B. DATA PREPROCESSING

The tweets we had collected, had too many unnecessary information, which we don't require for our system. For example, we don't require tweet's id or tweet creation time for recommending as artist. Therefore, we have removed that extra information. After that, we found that there were many redundant tweets in our collected tweets. We have also removed redundant tweets from the collected data set of tweets. After removing redundant tweets, we have removed following extra data from the tweets:

- Stop Words
- Emojis
- Slang Words
- Hashtag (#)
- Links

After Pre-processing phase finished, we were left with around 1000 tweets to use further for our system.

C. SENTIMENT ANAYSIS OF TWEETS

For finding the most popular or most loved artist, the important factor is the attitude/sentiment of the users while they tweet. Using sentiment analysis, we could know the user's sentiment towards the tweets is positive or negative. We have divided tweets as positive and negative and saved them into two different files.

After performing pre-processing and sentiment analysis, we have around 1000 tweets as dataset for further analysis.

D. DIVING TWEETS FOR TRAINING AND TESTING

In order to apply classifier, dataset should be divided into training and testing datasets. Most of the dataset should be used for the training datasets so that model can be trained more accurately.

For our system, we have used 60% of the collected tweets for training dataset and the rest for the testing.

E. LINEAR SVM (SUPPORT VECTOR MACHINE)

We have used Support Vector Machine classifier for our system. The important equation for understanding SVM is:

$$\vec{w} \cdot \vec{x} - b = 0$$

In the above equation, w is the normal vector to the hyper plane. The parameter b divided by w determines the offset of the hyperplane from the origin among the normal vector w. By SVM Model, the model accuracy is increasing highly. By using the linear SVM we have got many advantages to get success in the Project. Firstly, through the regularization parameter, user has avoided over-fitting. By using the kernel trick, we have built in expert knowledge about the problem via kernel and engineering. SVM is introduced by the convex optimization problem and for that there are so many efficient methods. Support Vector Machines are known as Supervised Learning Models in the machine learning with associated learning algorithms which analyzes the data used for classification and regression analysis, in these, for our project it was used for classification to differentiate the unique tweets into the positive and negative tweets. In our project, firstly the tweets were not labeled. So, in the beginning we have labeled the tweets and then we have applied the supervised learning algorithm for upcoming new tweets.

F. APPLICATION OF SVM ON TWITTER DATASET

The first task for applying classifier was to find the most frequent 10 features. To find the features, we have removed following from the tweets:

- Trivial words
- Spaces
- Alpha numeric characters

We labeled tweets manually as positive or negative. After removing above mentioned words, we found top 10 features and applied SVM classifier. After applying SVM classifier and 10-fold cross validation, we got predicted labels for the unlabeled tweets. The following figure shows the accuracy we achieved:

```
0.66
[0.78 0.62 0.68 0.74 0.86 0.78 0.54 0.68 0.42 0.52]
Accuracy: 0.66 (+/- 0.26)
```

G. ASSOICATION RULE MINING

We have generated Word Cloud for our collected tweets.

Association rule is useful for predicting and analyzing the user's behavior. Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness.

```
User 429: singer,rydell,siriusxm
0.0146 0.0588  singer -> renowned
0.0063 0.0252  singer -> jack
0.0104 0.0420  singer -> sarah

User 430: renowned,singer
0.0063 0.0252  singer -> jack
0.0104 0.0420  singer -> sarah

User 431: singer
0.0146 0.0588  singer -> renowned
0.0063 0.0252  singer -> jack
0.0104 0.0420  singer -> sarah
```

We have used python script to get association rules and generated one file having association rule, support value and confidence value. For finding the most popular artist/band to the specific location, support value plays a vital role.

Support of a rule is a measure of how frequently the items involved in it occur together. Therefore, if an Association rule has artist/band name and the highest Support value, then that artist/band is the most popular to the specific location.

V. RESULTS AND ANALYSIS

As mentioned above, we have collected tweets of Albany, New York and processed all of the above-mentioned steps and as a result we have got top 3 most popular artist of the Albany:

1. Sarah Hyland – Singer and Actress from New York City

2. MCR (My Chemical Romance) – Rock Band from Newark, New Jersey

3. Jack & Jack – Pop Duo from Omaha, Nebraska

These artists have the most Support value amongst all the association rules we got and thus they are the most popular artists of Albany city according to our system.

VI. LIMITATIONS

Our recommendation system has the following limitations:

- We have collected tweets using Twitter's REST API and due to limitation of API, we have collected comparatively small number of tweets and therefore the results we have got, could have been more accurate if we have more tweets.
- We have selected Albany region and found that there is lack of music related tweets.

VII. FUTURE SCOPE

For the future scope, we will try to collect more number of tweets so that we could get more accurate results and benefit local event management companies.

Current system is just for Albany, New York location, but for future scope we will try to include more locations. In fact, we will try to design our system in such a way where user needs to just enter the location and the most 5 popular artists/bands would be shown as a result.

We will try to design Graphical User Interface(GUI), so everyone can use it without having knowledge of python or data mining concepts. User will just have to select region or

location from GUI and then press enter and he will get the most popular artist/band/genre as output.

VIII. CONCLUSION

We came up with the problem statement. Many event management companies were facing this challenge, as there are many recommendations available which recommend an artist/band according to user ratings, but to find popularity of the artists based on location, there were no reliable system. The same problem was with the artists, because if they know that the people of specific location loves what kind of genre then it would get easy for them to perform or release new albums/songs for the audience of that specific region. This paper presented an approach by which we could easily find the most popular artists/ band / genre to the specific location. Therefore, we selected Albany region and collected tweets of Albany using music related words as a query. After Performing Data Pre-processing, Classification using SVM classifier, 10-fold cross validation and Association Rule Mining, we got the most popular 3 artists/bands.

IX. REFERENCES

[1] Personalized Music Recommendation by Mining Social Media Tags (https://www.sciencedirect.com/science/article/pii/S1877050913009009)

[2] Music-Map: A Music Recommendation System Based on Data Mining Technique (http://www.iosrjournals.org/iosr-jce/papers/Vol16-issue2/Version-11/R01621195100.pdf)

[3] Twitter Statistics. Available: (https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/)

[4] Kecman, Vojislav; Learning and Soft Computing — Support Vector Machines, Neural Networks, Fuzzy Logic Systems, The MIT Press, Cambridge, MA, 2001. Available: http://www.support-vector.ws/

[5] https://www.slideshare.net/dalmiaayushi/

[6] M. Jalali, N. Mustapha, M. Sulaiman, A. Mamay WEBPUM: a web-based recommendation system to predict user future movement

[7] https://www.techopedia.com/definition/14650/data-preprocessing

[8] https://ieeexplore.ieee.org/document/1706489/