Assignment 3 - TDT4173

Stian Ismar

1 Introduction

Millions of people use music streaming services every day. In the paper *Music Recommendation: Audio Neighbourhoods to Discover Music in the Long Tail*, the authors are presenting new ways to recommend music for users. S. Crew, B. Horsburgh and S. Massie want to move tracks out of the *long tail*. On an online music streaming platform, the long tail consists of tracks with fewer *tags* and plays.

Traditionally, collaborative filtering by using tags has been a way for proposing new songs for the user of a online music service. Songs with similar tags, i.e "rock" or "pop", can this way be compared and recommended to the user.

Collaborative filtering works great if a track has many tags. However, how can a song be recommended if it has little to no tags? Two new recommenders are presented in the paper. The new techniques utilise collaborative filtering, as well as the *sound representation* of the song. These two recommending techniques are tested on a large data set in a user experiment.

In this essay, I will present the research goal of the paper, the research methodology and the results. I will also discuss how the different parts were presented to the reader.

2 Research Goals

The main goal of the study behind *Music Recommendation: Audio Neighbourhoods to Discover Music in the Long Tail* is to improve the way music is recommended. The authors want to examine new ways of doing this. In addition, they want to compare the different methods up against each other in order to find the one that maximises user satisfaction.

3 Research methodology

In this section, I will present prior knowledge from the paper, the new recommenders and how they were evaluated experimentally.

3.1 Prior knowledge

The paper states that there are primarily three ways song recommendations can be made. One can use:

1. Tags: title, year, artist
2. Audio representations: texture, harmony, rhythm
3. Semantic information such as social tagging (collaborative filtering)

By using only social tags from collaborative filtering to recommend songs, one could experience that the less popular songs would diminish in relevance. The tags would be unevenly distributed and new and niche songs would continue to stay in the Long Tail.

One other way of recommending tracks is to use audio representations. All songs will to some extent have a texture or rhythm. However, if a song does not have a sufficient audio representation, a system only relying on the audio would be useless.

Auto-tagging is a way to add tags to songs. Auto-tagging is social tagging done by computers. The advantage is that this method gives most songs tags. On the other hand, this may lead to unfitting tags.

3.2 New recommendation techniques

Crew, Horsburgh and Massie were intrigued by the idea of combining some of the above techniques to make a better system for recommending songs. They made two new recommenders which would reduce the semantic gap between audio content and tags.

The two techniques they implemented were:

1. Learning Pseudo-Tags
2. Augmenting tags with Pseudo-Tags

The first one, *Learning Pseudo-Tags*, gets tags from a similar song and transfers them. The first step is that it finds the k most similar songs in regards to the audio texture. It then uses a rank-based sum of the tags in the neighbouring songs to return a pseudo-tag vector with pseudo-tags. This algorithm evens out the tagging distribution.

The pseudo-tag method is not optimal if a song already has a lot of tags. It does not take into account the other tags already on that particular song. *Augmenting tags with Pseudo-Tags* however, merges the pseudo-tag vector with the existing track tags. If the track has a lot of associated tags, fewer of the pseudo-tags are merged. If it is sparsely tagged, more of the pseudo-tags are added.

3.3 The data set

The data set called "The Million Song Data set" included a tagged data set with 950.000 songs with 500.000 unique tags. 46 % of the songs in the data set did not have any tags.

3.4 User evaluation

The research goal was addressed by doing a user evaluation of three different methods for recommending music. The three different methods were the Pseudo-tag, Augmenting tags with pseudo-tags (hybrid) and the traditional tag-method. The goal of this user test was to evaluate the quality of the recommendations and to see how many new discoveries were made. These methods were tested on the large last.fm data set.

The testing was designed to avoid bias. For every test, the user was shown the top five recommended tracks from one of the methods. Each of the shown songs had information and a 30 second mid-track sample. The recommender method is chosen randomly, and the song recommendations are also in a random order.

The user gives feedback on the quality of each recommendation. In addition, the user can select if they know the artist and/or song from before.

132 users participated in the test and over 1444 answers were made. Each user was also classified from a series of questions, in regards to music interests and so on.

In addition to the user experiment, a system centric evaluation was also performed using the last.fm user data.

4 Result

The quality of each recommendation is calculated by aggregating the individual scores by each user for that recommender. The results show that the hybrid recommender provides a better recommendation, followed by the tag-only recommender and the Pseudo-tag.

S. Crew, B. Horsburgh and S. Massie found out that users who previously knew of a track or artist, would give that recommendation a higher quality rating. Interestingly, one can see from the results presented that the hybrid recommender received the highest rating for all types of songs- known artist, known track and neither (novelty).

When novelty % is plotted against score quality, we see from their results that the hybrid gives the best results.

The system centric evaluation found the same results as the user experience. The hybrid and tag methods are similar, with the hybrid coming out on top. The hybrid performed better when the number of recommendations N is larger than 5.

5 Evaluation

For the user experiment, the results of the user tests are evaluated by summing up the scores for the different recommenders. For the system centric evaluation, a function called socialSim was developed to set the recommendation quality. SocialSim uses the association between the numbers liking and listening to tracks q and r.

I think their justification of the result is thorough. During the testing, they randomised as much as possible, which makes the paper credible. The evaluation methods are well presented, which also makes it clear how they analysed the results.

The authors chose to use a data set with almost 50 % of the songs untagged. This way, they could really test the different methods on novelty tracks, which was the main goal.

6 Discussion and own thoughts

Strengths and weaknesses for the recommendation techniques are portrayed in the paper. For example, a strength related to using audio representations is brought up, since it allows new songs to get content tags. On that same topic, negative sides are addressed for systems relying solely on the quality of the audio for recommendation. The same goes for when the social auto-tagging is discussed.

I like the way the paper i built up because it is easy for the reader to follow the development of the hybrid system. At first, the different track attributes are presented, and then the prior ways of recommending new songs. As I continue to read, more and more information is added to make the recommenders more and more complex. They also build on each other. The new hybrid recommender uses the power of the pseudo-tags from neighbouring tracks, and merges the tags to the particular track.

The different recommenders are also discussed in the conclusion. For instance, they write that recommendations based on social tags are influenced by others. That makes sense, because a user can misplace a tag or have another opinion of what kind of tags a track should have. Solely tag-based systems are also bad for moving tracks out of the long-tail.

In section 5.3, the authors evaluate their results from the user experiment. They are unsure of why the top recommendations by pseudo-tag is poorer than those that are ranked lower. I think it is good that they address the uncertainty here. It gives the paper credibility. They also present a possible reason as to why this is.

In the user experiment, I think it is great that they made the user answer questions regarding familiarity to the recommendations. Consequently, they had more data to study and could see the increase in quality for the hybrid recommender on the quality-to-novelty plot.

With the increase in popularity for online music streaming, I think this paper played an important role when it was released in 2015.

Lastly, I want to say that the presented research is important and positive for the music industry. With the hybrid recommender, users get better song recommendations. Additionally, less know artist will have a better chance of being listened to, which I think is good. Since I use music streaming and use music recommendations daily, I thought the paper was very interesting. It made me more aware of all the work that is behind the song recommendations.

7 Sources

http://www.businessofapps.com/data/spotify-statistics/

Music Recommendation: Audio Neighbourhoods to Discover Music in the Long Tail- Susan Craw, Ben Horsburgh, and Stewart Massie