ASSIGNMENT 3

TDT4173 - Machine Learning and Case-Based Reasoning

Written by:

Håkon Ødegård Løvdal

March 2016

NTNU

Norwegian University of Science and Technology

1 Introduction

In the paper *Music Recommendation: Audio Neighbourhoods to Discover Music in the Long Tail* the authors Susan Craw et. al. [1] focuses on how to create a music recommender system, that includes music from the "long-tail". Long-tail music may often be just as relevant, as the mainstream music. Without certain insight and knowledge about music, this long-tail music if often hard to discover. This is the problem the authors of the paper wants to address. I feel it is safe to say that the authors manages to elaborate well about this problem. The paper follows an normal structure for academic papers to present and elaborate about the research. In 2015 the paper was awarded "Best Paper" at the International Conference on Case-Based Reasoning [2].

Initially an introduction to the concept of "long-tail" music and existing recommender systems is presented. In the next section (2), the authors introduce relevant literature concerning music recommendation and serendipity. Thereafter, in section 3, the music collection used during the research is presented. Details about the hybrid-recommender systems created during the research is then presented in section 4. As a natrual follow up to this, two different hybrid-recommenders created during the research are presented; *Pseudo-tag* and *Hybrid*. Pseudo-tag is a recommender that extracts tags from similar tracks in the audio neighbourhood. This is done by a k-nearest neighbours retrieval. The Hybrid uses tags that augments learned pseudo-tags with it's own tags. In section 5 and 6, two experiments conducted and their evaluation is presented. The two experiments conducted were respectively user evaluation by real users; and a evaluation using user-data from the Last.fm API. The paper ends with a conclusion, where the researchers elaborate about their findings.

2 Research goal

By my understanding, the main goal of the research can be summarized in one sentence found in the abstract of the paper; *"Users are looking for high quality recommendations, but also want to discover tracks and artists that they do not already know, new releases and the more niche music found in the "long tail" of online music"*. The authors argue that the current, traditional tag-based recommender systems are biased towards popular tracks, and therefore not efficient on long-tail music. In their attempt to address this research question, they present a fair amount of relevant work, which addresses similar issues. Examples of this is

papers on tags, hybrid-reccommenders, auto-tagging to mention a few. Of course the actual research questions were more precise than this, but I felt the above quote managed to encapsulate the concrete problem in a very satisfiable way.

2.1 Research methodology

The research conducted, were an experimental research. The experimentation were done conducting two experiments. Firstly, it were conducted an online study, using real users to test the quality of the recommendations on real users. Of course nothing beats real users when you are in search of this kind of feedback. Secondly, it were conducted a larger experiment using user-data from the Last.fm API. This experiment were performed on the entire test-dataset, and therefore it acted out as a support to the results from the first experiment.

2.2 Dataset

As stated before, the authors created a music collection as a dataset. The collection included 3174 tracks, from 764 unique artists. To be able to do the experiments with this collection, they applied two representations to the tracks. The first being a tag vector representing the track. This were based on tagging from Last.fm users. The second were a vector representing the texture of the audio. For the tag representation, a tag could be as simple as "jazz". As stated, the tags used were collected from the Last.fm API. Whenever a user on Last.fm listens to a track and tags it, this tag is strengthened (counter incremented). Each track's tags is represented with a tag vector $t =< t_1\ t_2\ \ldots\ t_m\ >$, where t_i is the frequency for a tag i. Totally, the collection contained 5160 tags.

The texture representation, were represented using MFS Mel-Frequency Spectrum. This were done by first splitting the audio waveforms into windows of 186 ms, and then convert each of these windows into a frequency domain using Discrete Fourier Transform. Thereafter, each window is changed to a feature vector, using the mel-scale. A mean feature vector MFS is then computed for the track. Also Latent Semantic Indexing is used to capture musical texture concepts. Together this yields the MFS-LSI texture vector representing the track. [1]

[1]The two previous paragraphs is in many ways almost a pure transcription of the same paragraphs in the actual paper, but it were hard for me to present it in another way due to lack of knowledge regarding this methods.

3 Results

For the experiments, two hybrid recommendation systems were created. The two recommenders main goal were to reduce the semantic gap between the audio content and the tag vectors of tracks. According to the paper a general hybrid recommender is a system that merges multiple representations of tags and audio, or it combine processes of subrecommenders. The authors claims to have created a new and different approach to this. Their system is said to exploit the similarity assumption of case-based reasoning to extract additional tags from audio neighbourhoods of tracks. With this approach, they are able to augment existing tags of tracks to give similar tracks useful and meaningful tagging.

Also the recommenders were intended to increase recommendation quality on less tagged tracks. The first of the created systems were the *Pseudo-tag* system. In this system, tags are extracted from tags of similar tracks (using audio/texture as the measure). Similar tracks are found by using the k-nearest-neighbour algorithm with cosine similarity on the MFS-LSI texture vectors. The authors argue that the approach using pseudo-tags is better than audio content directly, since context and opinions of users will be inherited from the neighbouring music. In my opinion this argument seems valid, and I think it will produce some interesting results.

The second of the created systems were the *Hybrid* system. Here a tag representation is generated by merging the learned vector of pseudo-tags, p, from the Pseudo-Tag system with a tracks tag vector, t. According to the authors, this approach solves a problem that pseudo-tag vectors alone is useful when a track has few tags, but is to influential on the result when a track is well-tagged.

3.1 User Evaluation

The first experiment conducted were a user evaluation on real users. I am not going to elaborate about the details of the experiment, but focus on the results. The experiment were conducted online and lasted for 30 days. 132 users submitted their results. This summed up to a total of 1444 queries evaluated, but 386 of them were discarded, due to an assumption that the users had clicked without adjusting any scoring parameters.

The result of the recommendation quality were calculated by aggregating the individual scores for every user, across all users U that provided a feedback for the recommendations. A formula $Q@N$ is presented to give an average of the top N recommendations to evaluate the recommendations for a query, q. In general the

Hybrid recommender yields better recommendations than both Tag and Pseudo-tag.

An assumption presented by the authors is that users give higher feedback to known music. This makes sense, does it not? The results for the recommendations of new and niche tracks shows that users gave a higher feedback to known artists and tracks. The less information available to the users, the less quality of the feedback. This proves the presented assumption. Although, at the same time, the experiment shows that the Hybrid recommender gives a higher quality on the recommended tracks, even though the user did not know the artist or track on beforehand (the users' bias on known artist and tracks).

3.2 Last.fm User Data

The second experiment conducted were a large test on the entire music collection. User data from Last.fm were used to get how many likes and listens a track had. Then this data were entered into a formula to calculate the quality of the recommendation. In order to evaluate the quality, the authors developed a similarity score, $socialSim$, representing the association between likers and listeners of a track q and the recommendations r. The results from this experiment confirms the results from the user evaluation. Therefore it supports the results from the first experiment. The Hybrid recommender seems to provide recommendations with an overall higher quality than Tag and Pseudo-tag.

4 Evaluation

The results of both experiments seems to prove that the Hybrid recommender provides improved and good recommendations, but are the results evaluated in an acceptable way? I would say so. Both the $Q@N$ formula for evaluation of recommendation quality, and $socialSim$ for evaluation of Last.fm user data is described in such a way that I feel confident that they provide reasonable and legitimate results.

The only thing I somewhat react to, is the size of the user participation experiment. Even though the data seems to yield a positive result for the research, I think that 132 users is a rather small amount of people. Is this amount representative enough to state that the recommender-system in general provides good recommendations? I would say no.

One thing I feel could be discussed more thorough, is known or possible weaknesses of the recommender system. I can't find any clear discussion on this when reading the paper. The only place I find a clear indication of a possible weakness being discussed is during the evaluation of the Last.fm User Data experiment.

All in all, when I were done reading the paper the first time, I felt confident that the authors had solved the question in mind in a satisfiable way. They wanted to create a hybrid-recommender that bridged the semantic gap between tagging and audio. I felt the data provided justified the results, and it made me confident that they had created a successful recommender.

5 Did I like it?

I really liked the paper. It manages to give a detailed and thorough elaboration about the theory, research, related work and results in a way that is easy to understand and follow. The structure of the paper is very good. Especially the evaluation criteria I find descriptive. The part I find hardest to understand would be the part regarding how the authors designed their recommender-systems and representation of tags and audio content. This is probably due to that the paragraph about the texture representation of audio files contains a fair amount of technology and concepts I have little to no knowledge about.

References

[1] S. CRAW, B. HORSBURGH, AND S. MASSIE, *Music Recommen- dation: Audio Neighbourhoods to Discover Music in the Long Tail*, Case-Based Reasoning Research and Development, (2015), pp. 73–87.

[2] A. DUTHIE, *Rgu researchers transform online music discovery*, 2015. URL: `http://www.rgu.ac.uk/news/rgu-researchers-transform-online-music-discovery/`.