Musical Affective Recommender System (MARS)

Daphne, Rens, Everton, Armen, Jesse

Abstract—A Music Recommender System is introduced, founded on text-based sentiment analysis and questionnaires to determine the affective load of music signal characteristics, and developed and validated with data from the Million Song Dataset and EchoNest public API. [Conclude with our findings].

Index Terms—Computing science, sentiment analysis, music recommendation, paper

1 Introduction

MUSIC preference has many dimensions, one of which is the emotional state of the listener [1]. It is difficult to imagine listening to loud, high-energy music when one wishes to remain calm or focus on a task, as suggested by J.H. Janseen et al [2]. Additionally, music can have a therapeutic effect on the listener. For instance, studies have taken this approach to use music to aid cancer patients and their families [3]. In our project, we are interested in the relationship between the listener's mood and music. We aim to identify the listener's emotional state by analyzing a text description, and then provide songs that can either match their emotional state or improve it.

For the first part of our project, we performed a survey to identify well-established techniques and the state of the art. We begin by researching ways to quantify emotion. There are different models used for identifying the specific emotional state which is closest to a song. Existing techniques on emotional analysis based on music extract properties from a song like BPM, genre and timbre [4]. After extracted from the song, these features are used to quantify the mood or emotional context. For the following part of our project we do the inverse mapping. We identify ways to connect the listener's mood to a song. In Tune into your emotions (Janssen et al. 2011) [3] it is described how biosignals are translated to emotions and music is retrieved to affect these emotions in a specific way. Our project is similar, however, our approach will be more business oriented and we will be gathering emotions in a more accessible way; from text. In this phase, we will research how accurately we can give music suggestions based on emotional analysis of text fragments.

After researching these topics, we aim to produce a multimedia system that is able to identify the listener's emotions through text and map them to songs that match this mood. As suggested by Sauro & Lewis [5], the main factor of a success for a system is if it can capture the user as the main actor. We aim to achieve this by providing an intuitive user interface and being mindful of the user when constructing our algorithm. During this process, we found novel ways to use emoticons for this purpose.

Lastly, this task presents valuable societal impact and business potential. The societal impact of recommendation systems is great for it allows for media discovery, allowing unheard artists to be heard. Additionally, with the identification of emotional content we bring a new dimension to recommendation systems in general that can better gauge the receptiveness of a user to media. This gives an opportunity to more advertisement aware platforms. Through this project, we hope to both explore the state of the art of multimedia technologies whilst also producing a robust multimedia system.

mds
August 26, 2015

2 Survey

In this section, we describe the prevalent approaches in sentiment analysis. The simplest one only distinguishes positive and negative sentiments [4]. Other categorical representations are the 66 adjectives arranged in 8 groups as identified by Hevner (1936) [4]; the 146 emotional terms specific for music mood rating proposed by Zenter et al.[4]; the 5 mood clusters from labels from the All Music Guide, as used in the MIREX competition; and PASAS. [4].

Source [4]

Another approach is to map emotions to a dimensional model. The most common dimensional model is the valence-arousal space, as proposed by Russel and Thayer [9]. It is a 2D model, with valence on the x-axis and arousal on the y-axis. Valence tells something about the positivity of an emotion, so happy has a high valence, but the valence of miserable is very low. Arousal is about the excitement that is attached to an emotion, so alarmed has a way higher arousal value than tired.

In our system, we choose the valence-arousal model. With this model, we expect to be able to suggest songs of which the emotion fits relatively well to the emotion in the text that was inputted by the user. If we would choose a categorical model, every song in a category would have the same probability of being suggested - this might give the user the impression that the music was chosen randomly and not based on his emotion. Using the dimensional valence-arousal model, we can easily compare distances between the emotion in the text and in the music, compare

distances and suggest the song to which the distance is smallest.

3 METHODS

3.1 Text Analysis

The text sentiment analysis is done as follows. We map values of the text given by the user to a valence-arousal space, by a combination of four approaches: evaluating text, punctuation, textual emoticons, and emoticons. We used a dictionary that maps English words to corresponding valence-arousal values, this is the same approach used by A. B. Warriner et al [3]. This approach computes the valence-arousal by using bag-of-words model. The texts grammar, order and context are disregarded.

Figure 2. Questionnaire.

However, we have expanded this approach to include punctuation and emoticons. This modification is done to improve the valence-arousal determination on short text. This is because sentiment analysis on short text is challenging and does not present satisfactory results or has limited categorization of emotions. As such, the valence-arousal values for punctuation and the two kinds of emoticons were determined by the questionnaire [Fig. 2], and handed out to 50 people to fill in. We will use certain weights for the keywords, the emoticons and the punctuation to determine the total valence/arousal values of the user given text.

3.1.1 Audio Analysis

In order to obtain the emotion values (valence and arousal) for each song in the database, we trained a music emotion regression model. In this section, well explain how this was accomplished. Due to limited time and scope of this projected we sought pre-collected music metadata. Our requirements for this data were; a broad range of genres, audio features, and lyrics. The Million Song Dataset [4] was the best candidate and the dataset used in our system. The biggest advantage is that this dataset is the largest one that is freely available at present. It consists of a million contemporary popular music tracks, so the music would be appealing to a great audience. Moreover, each song comes with audio features and metadata.

The Million Song Dataset consists of one million audio features and metadata belonging to contemporary popular music tracks, in HDF5-format. These are provided by The Echo Nest, a music intelligence company, which was recently (March 2014) bought by Spotify. A list of all fields of files available in the dataset is listed on http://labrosa.ee.columbia.edu/millionsong/pages/field-list. There is a lot of information available for each song (all kinds of metadata, corresponding IDs of this song in other datasets, etc), but we are mainly interested in audio features like loudness, mode and energy. Even more audio features can be obtained by using the Echo Nest API. Additionally, Echo Nest offers a psychological label; valence. Using a combination of the valence and energy features, we can easily map each song to valence-arousal space. Note however that this only takes audio the information into account and does not use other data that could be relevant for the emotion associated with a song, for example, lyrics or tags. For our project we used 4006 songs from the Million Song dataset that possess valence-arousal provided by EchoNest. You can observe the valence-arousal space of this data on figure 3.

Figure 3: Valence and arousal values for 4006 songs. These values were retrieved from the EchoNest.

Additionally, the Million Song Dataset does not offer lyrics data. Fortunately, a connection to other datasets that do can easily be made, as the Million Song Dataset provides IDs of a song in those other datasets. Lyrics in bag-of-words format are provided by musiXmatch and are directly associated with 237.662 MSD tracks. Tags can be obtained from the Last.fm dataset - 505.216 of the songs from the Million Song Dataset have at least one tag. Audio data is not in the MSD for copyright reasons.

Furthermore, in our Music Emotion Recognition component, we combine the emotion obtained from the audio data with the emotion obtained from the lyrics. Research by Yang et al [4] shows that this multimodal approach can greatly enhance audio-based classification algorithms: a relative improvement gain in classification accuracy of up to 21

There are various possibilities for doing the fusion of audio and lyrics. Yang et al. [4] mention three multimodal fusion methods: Early-fusion-by-feature-concatenation (EFFC): combine feature vectors of audio and lyrics to a single feature vector and make a classifier from this. Late-fusion-by-linear-combination (LFLC): train classifiers from audio and lyrics separately and combine their predictions afterwards in a linear way. Late-fusion-by-subtask-merging (LFSM): train audio and lyrics classifiers to classify valence and arousal separately and merge the result afterwards. So the difference is that the two dimensions go separated into the merging step.

Table 4: Performance comparision of a number of multimodal fusion methods for four-class emotion classification, arousal classification and valance classification, from [8]

The researchers compared the performances of these approaches and found out that LFSM (see table 4 for results) gave the best accuracy results, as can be seen in Table 4. However, we see that the accuracy for valence is higher when using LFLC.

That is why we choose a combination of LFSM and LFSM in our system, which is possible as we do regression. We will calculate the valence as a linear combination of the valence from audio features and lyrics, while we calculate arousal by using just the audio features:

Valence(audio, lyrics) = Valence(audio) + (1-) Valence(lyrics) Equation 1: Formula for determing valence of a song, using the valence value from both audio and lyrics. is a scalar value between 0 and 1.

Arousal(audio, lyrics) = Arousal(audio) Equation 2: Formula for determinging arousal of a song, using only the arousal value from audio.

Initially, we will choose 0.5 as value for . In the experiments, we will also experiment with other values.

3.2 Emotion from lyrics

As explained previously, the lyrics we retrieve from musiXmatch are in the bag-of-words model. This limits our possibilities for the lyric emotion detection algorithm a little,

as we can only use unigrams. Our plan is to adapt the sentiment analysis algorithm we already created for the user input to lyrics. There are some differences: we do not interpret emoticons and interpunction in lyrics, as these do not occur in this domain. Also, we can omit the step of converting the text to the bag-of-words model, because the lyrics are already in this format. The motivation behind our approach is that in this manner we can best match the user's mood during sentiment analysis.

3.3 Reverse engineering of audio energy and arousal

The valence and arousal values we use from the audio are taken almost directly from the Echo Nest API. Unfortunately, little is provided in regards to how they obtained these values. However, we expect them to perform reasonably well (at least better than we could do in the scope of this project), but still we would like to know on which features their algorithm is based. If we know something about this, then it is easier to create our own algorithm so we would be less dependent on a company.

That is why we extracted several audio features from the Million Song Dataset. We chose features that are relevant for Music Emotion Recognition: loudness, tempo, mode, mode confidence, key, key confidence and timbral features. The latter are MFCC-like features, which are 12-dimensional vectors with timbral information, calculated for each 30-second segment of a song. We calculate the mean and standard deviance for each element over all the segments, so we get a 24-dimensional vector for timbre of a song. Using the valence and energy labels from the Echos Nest as ground truth, we then trained a regressor on part of the data (the training set) and tested its accuracy on the other part (the test set).

For some songs, the mode confidence and/or key confidence is very low. In this case, we dont want our regressor to take mode and/or key into account for that song. That is why we replaced the mode value from songs with a mode confidence below 0.25 by the mean mode value. We also replaced the key value from songs with a key confidence below 0.25 by the mean key value.

For performance evaluation, we used the R2 statistic, which is calculated as follows: R2(y, r(X)) = cov(y, r(X))2var(y) var(r(X)) Equation 3: the R2-value when comparing the observed value y to the predicted value r(X).

We splitted the data into a training set of 2000 songs and a test set of 1992 songs. Then we tried some regression methods: multiple linear regression and support vector machines with a linear, polynomial and radial kernel. Each time, we trained a regressor for valence and arousal separately. We report the resulting R2-values for the training and test set in Table 5.

Table 5: R2-values for regressors on training and test set

We can conclude that the SVM regressor with a radial kernel gives us the best model. However, the results are not that good so to get a better fit, we probably need to take other audio features into account. In future work, we could do more experiments with also for example pitch features.

4 System description

In this section we give a broad overview of our application as well as describe how our work can be reproduced. The description offered here is a user centric description, so we first start with the user interaction with the system. The user main interaction is with a text box. In this text box the user enters text about their current mood.

Sentiment analysis is done on the text inputted. As previously mentioned in our methods section, we use a dictionary approach to map each word entered to a valence and arousal value. These values are matches to the closest valence and arousal values present in the database. A song is then selected. Finally, youtube is searched by using the song title and artist name. This video is then played to the user.

Figure 4: System overview of MARS.

5 System Performance

Results; that is, system performance, from two perspectives

6 Discussion

There are some limitations in our project. In this section we discuss them and suggest solutions that can potentially mitigate these issues. One of the main factors we omitted in our project is music preference. As suggested by Janssen et al [1] music preference changes how the user is affected by music. The example given is that listeners familiar with high tempo music such as heavy metal may perceive Heavy Metal songs to be less arousing than someone not as familiar with the genre. Due to the scope and time we were unable to capture this dimension in our system. However, given the plethora availability of music services online, this task is not unfeasable. This issue could have been mitigated by requested user data from services such as Spotify. This way we could better determine what the listener listens to most often and their genre preferences.

An additional limitation with our approach is the data used, specifically, the valence-arousal values retrieved from Echo Nest. Little information is provided on how these values are produced. Nevertheless, there are reasons to believe that these values are not labels produced by humans but rather the result of a proprietary algorithm. Due to the time and scope of our project this is an issue that is difficult to mitigate. However, observation of samples of the dataset indicate that the values matches what a listener would expect. We also attempted to identify and reverse engineer the process using a regressor model. Due to our limitations however it was not possible to achieve ground truth with the dataset provided. We strongly recommend further work to do this as it may improve results.

7 Conclusion

The conclusion goes here.

Appendix A
Proof of the First Zonklar Equation

Appendix one text goes here.

Appendix B

Appendix two text goes here.

ACKNOWLEDGMENTS

The authors would like to thank...

REFERENCES

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

Michael Shell Biography text here.

John Doe Biography text here.

Jane Doe Biography text here.