A Probabilistic Topic Models Based Music Recommendation System

Lixuan Zhu
Rutgers University-New Brunswick
lz306@scarletmail.rutgers.edu

Yu Zheng
zy120@scarletmail.rutgers.edu

Yi Zhong
yz614@scarletmail.rutgers.edu

Shihao Su
ss2719@scarletmail.rutgers.edu

Abstract

We will implement a music recommendation system utilizing Probabilistic Topic Models. Specifically, given a song, the system will find a series of similar songs according to features such as metadata, tags and acoustic characteristics.

1. Introduction

Topic Models are a collection of algorithms that finds the underlying topics from a large amount of text documents. According to Blei, each document is a mixture of corpus-wide topics; each topic is a distribution over words and each word is drawn from one of those topics [1]. When a distribution of topics of the document emerge from topic modeling, we can compute the similarity between two documents by comparing their topic distributions, thus the documents with high similarity to the given document can be recommended to user.

There are two general approaches when designing recommendation systems: collaborative filtering and content-based filtering, each with their strengths and weaknesses. Collaborative filtering utilizes the preferences of existing users and predict whether a specific user will like an item based on the decision of similar users. Music streaming services such as Last.fm utilizes collaborative filtering to recommend songs based on specific user profiles [4]. However, a natural challenge for collaborative filtering methods is the lack of user ratings, also referred to as the cold start problem [4]. For example, when a new song is published, there is not enough user rating for the system to know which users will like the song. The other method, content-based filtering, focuses on the similarity of the songs itself. Pandora, another popular music streaming service, utilized the metadata(artist, genre, etc.) and tags to find similar songs given an initial seed song [3]. Although content-based approaches require very little initial information, it is limited to recommend the songs similar to the original seed since the information about the song is limited.

In this project, we take the content-based filtering a step forward. We utilize two probabilistic topic models, latent dirichlet allocation(LDA) and hierarchical dirichlet process(HDP), to discover the similarity between songs and compare their performances. We consider each song as a document and the features of the song as words. It is important to define what we mean by features. Traditional recommendation systems uses the metadata of the song itself and tags by users as measures for similarity. In our model, we incorporate acoustic properties of the music such as pitch, tempo, verses and chords used, mood progression, etc. The timeline API from Gracenote is used to retrieve acoustic properties from the songs[2]. Collectively, we consider the traditional measures and acoustic properties of the song as features. Thus each feature is a word in the document. The probabilistic topic models process the entire collection of songs as a corpus. The advantage of utilizing acoustic properties is the range of songs recommended can be extended. Songs with similar acoustic properties may not be in the same genre or tagged with similar attributes. The details of the implementation of our models and the challenges of utilizing acoustic properties will be covered in section 3.

Another focus of our project is to compare two topic modeling algorithms, LDA and HDP. We plan to analyze the performance of both algorithms in terms of complexity and recommendation result. In comparison, our HDP model is based on LDA and more complex in terms of implementation. However, LDA requires a priori entry of the number of topics in the corpus. How to choose the input value and whether the recommendations are meaningful will be addressed in section 4.

2. Prior Work

3. Models, Assumptions and Requirements

3.1. Word Model of Songs

3.2. Latent Dirichlet Allocation

3.3. Hierarchical Dirichlet Process

4. Evaluation

5. Plan

The project will be finished by early December. The first step is to evaluate the existing implementation of LDA and HDP algorithms and see if any of them can be tweaked to fit our data. This step will be finished by Nov. 6th. The next step will be to explore ways to measure similarities between songs. When we have a distribution of features of the songs, how much weight do we put on each feature and how do we calculate the similarities? This step will be done by Nov. 13th. The third step will be building user interface to interactively make recommendations. This step will be done by Nov. 27th. The last step will be testing the system, evaluating the algorithms and writing project report. This step will be finished by Dec. 3th.

References

[1] D. M. Blei. Probabilistic topic models, 2011. https://www.cs.princeton.edu/ blei/kdd-tutorial.pdf. 1

[2] Y. Hu. *A Music Recommendation System Based on User Behaviors and Genre Classification*. PhD thesis. 1

[3] G. Inc. Entertainment metadata. https://developer.gracenote.com/metadata. 1

[4] N. R. M. Elahi, F. Ricci. A survey of active learning in collaborative filtering recommender systems. *Computer Science Review*, 20(1):29–50, 2016. 1