Xueli Fan. Track Co-occurrence Analysis of Users' Music Listening History. A Master's Paper for the M.S. in I.S degree. April, 2018. 40 pages. Advisor: Ryan B. Shaw

Music services provide listeners access to great numbers of available tracks. It is time consuming for listeners to find potential favorite ones. Music listeners increasingly want playlists to be created automatically. This study examines the relationship between background knowledge about music and track co-occurrence frequency in users' music listening history and builds a multiple linear regression model to predict the track co-occurrence. So given a seed track, the model can find out which track is most likely to co-occur. A simple objective evaluation compares predicted track with tracks in the users' listening history. 13 out of 15 test tracks find the highest rank predicted track in the same listening history.

Headings:

Music

Playlist generation

Co-occurrence

Correlation

Linear regression

TRACK CO-OCCURRENCE ANALYSIS OF USERS' MUSIC LISTENING HISTORY

by
Xueli Fan

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

April 2018

Approved by

_____

Ryan B. Shaw

Table of Contents

Introduction 2

Literature Review 5

2.1 Playlist 5

2.2 Playlist Generation Algorithms 6

2.3 Evaluation Approaches for Playlist Generation Algorithms 8

Methods 10

3.1 Data Collection 10

3.2 Data Analysis 12

Results 17

4.1 Correlation 18

4.2 Multiple Linear Regression 23

4.3 Evaluation 24

Discussion 27

Conclusion 33

Bibliography 35

Introduction

Our music listening habits have been changing dramatically in recent decades. Once people had to go to live shows, like concerts, to listen to their favorite music. Then technological innovation made it possible for audio to be recorded on a medium for playback, first on cylinders (Edison, 1878), then on acetate discs, cassette tapes, and CDs (Fields, 2011). But the number of songs recorded on a CD is still limited. Mobile technologies introduced MP3 devices to music listeners. Music listeners can have thousands of favorite songs in their devices and listen whenever and wherever they are. Since smartphones became widespread in the late 2000s, MP3 devices were gradually replaced by music player software installed in smartphones.

Whether they use MP3 devices or music player software, users can create playlists of their favorite songs. A playlist is defined as a set of tracks (audio recordings). Users can create playlists for different artists and genres or create a playlist for all the songs they like. A general playlist concept also covers other types of playlists, like album tracklists. Album tracklists are released by artists and labels and contain new tracks for one artist or one genre. In this paper, I focus on playlists that are made by listeners based on their listening history. The order of the tracks is also not considered in this paper, because music is widely played in a shuffle mode now and randomness can enrich user experiences (Leong, Howard, & Vetere, 2008).

Instead of creating playlists manually, music listeners increasingly want playlists to be created automatically. Music services provide listeners access to great numbers of available tracks. It is time consuming for listeners to find potential favorite ones. Listeners would like music services to help them make some decisions. This leads to the concept of playlist generation. “Given a pool of tracks, a background knowledge database and some target characteristics of the playlist, create a set of tracks fulfilling the target characteristics in the best possible way” (Bonnin & Jannach, 2015). The background knowledge includes all kinds of track information. Target characteristics means what characteristics users think the playlist should have.

Playlist generation is challenging. It is hard to find an objective way to evaluate the quality of playlists. Whether listeners like the playlist or not is very subjective. Track co-occurrence is a relatively objective and direct way to represent listeners' preferences. When there are tracks that frequently appear together in a collection of listening histories, the assumption can be made that the tracks both fulfill the preference (Baccigalupo & Plaza, 2006).

It is worth asking whether there is relationship between background knowledge about music and track co-occurrence frequency. Finding a relationship could help music services use background knowledge to predict track co-occurrence frequency and thus create high quality playlists.

In this paper, I investigate the following questions:
Is there relationship between background knowledge about music and track co-occurrence frequency?

What type of background knowledge about music best predicts track co-occurrence frequency?

Literature Review

This section introduces the history of the playlist and discusses some playlist generation algorithms and evaluation approaches proposed in previous studies.

2.1 Playlist

The birth of the playlist can be traced back to the 1850s, when the selection and ordering of pieces for concert programs began to be decided based on the ideas of a program director, instead of just maximizing coverage of various tastes (Weber, 2001).

Then radio and the phonograph were invented. These two technologies enabled the broadcasting of music to much larger audiences. This ended the need for the physical presence of the performing artists, which made the selection and ordering of pieces much easier. These are essential features for the playlist. The term “playlist” was first used to describe sets of songs when genres were promoted on the radio (Wall, 2007).

The playlist went personal as portable audio devices that played cassette tapes became available. Listeners could make their own mixtapes by selecting their favorite songs and ordering them (Bull, 2006). It was similar to how people create their personal playlists now. But with the internet and digital media storage, the pool of songs available to be added to a playlist increased a lot. With so many songs, it has been shown that it is less necessary for intentional ordering, as random order can lead to serendipity (Fields, 2011).

So, in this paper, I define a playlist as a coherent but unordered set of songs, made by listeners based on their personal preferences.

2.2 Playlist Generation Algorithms

In the literature, there are some common playlist generation strategies: similarity-based algorithms, collaborative filtering, frequent pattern mining, case-based reasoning, context-awareness, and hybrid strategies.

It is easy to think of selecting tracks based on their similarity to generate playlists. Pampalk, Pohle, & Widmer (2005) tried to generate playlists automatically based on skipping behavior and audio-based similarity. The evaluation showed that the generation reduced the number of tracks skipped. Flexer, Schnitzer, Gasser, & Widmer (2008) proposed an approach which was based on audio similarity and did not require any kind of meta-data. They created a smooth transition through tracks based on a start song and an end song. However, pure similarity-based approaches have the problem that the generated playlists are too homogeneous and thus not satisfying. Listeners do not want too many tracks from the same artist. One of the reasons they use automatic playlist generation is to discover new tracks.

Collaborative Filtering (CF) is another playlist generation strategy. It is the prevalent approach used in the field of Recommender Systems (RS) (Jannach, Zanker, Ge, & Gröning, 2012). CF approaches make predictions about user's rating for an item, based on the existing ratings of other users who have ratings similar to those of the active user. Music recommendation is a special case of the more general recommendation problem, so CF approaches can also be applied to playlist generation. Track rating data can be

obtained explicitly or inferred through analysis of listening logs. Chen, Moore, Turnbull, and Joachims (2012) presented a playlist prediction approach that is analogous to matrix decomposition methods in collaborative filtering. CF approaches have the limitation that they cannot be used for new users without any rating information.

Frequent pattern mining approaches rely on the co-occurrence of items in playlists. These approaches try to identify global patterns in the playlist data. Patterns include association rules (AR) (Agrawal, Imieliński, & Swami, 1993) and sequential patterns (SP) (Agrawal & Srikant, 1995). The order of tracks is considered in SP, while it is not in AR. Frequent pattern mining is one of the most straight-forward solutions to playlist generation. According to Bonnin and Jannach (2013), a number of algorithms for efficient pattern mining work comparably well. However, they are not frequently used in the literature.

In the Case-Based Reasoning (CBR) process, every playlist is seen as a case whose relevance is inferred measuring the co-occurrences of its songs in a large collection of past playlists. Baccigalupo and Plaza (2006) presented a CBR approach to musical playlist recommendation. They introduced a ‘knowledge-light’ approach to recommendation, based only on user-related knowledge.

Context-aware music recommender systems can be contrasted with content-based systems. The context of listening can influence and be influenced by the music and the listener. Hariri, Mobasher, & Burke (2012) presented a context-aware music recommender system which inferred contextual information based on the most recent sequence of songs liked by the user. Their experimental evaluation showed that their system could give better recommendations than a conventional recommender system

based on collaborative or content-based filtering. Vigliensoni and Fujinaga (2014) tried to identify the time zone where listeners were by analyzing listening logs. They suggested that the location of listeners can be an important contextual dimension used in context-aware music recommendation systems.

Hybrid strategies combine different playlist generation techniques. Hornung et al. (2013) presented a weighted hybrid recommender approach that amalgamated three diverse recommender techniques into one comprehensive score.

2.3 Evaluation Approaches for Playlist Generation Algorithms

The evaluation of playlist generation is difficult, because deciding whether a playlist is satisfying or not is very subjective. Researchers have done some surveys on this tough topic (Fields, 2011; McFee & Lanckriet, 2011; Bonnin & Jannach, 2015). I divide evaluation approaches in two general categories: subjective evaluation and objective evaluation.

Subjective Evaluation

Since playlist generation is a music recommendation and discovery service for the user, user satisfaction is the ultimate goal. The most direct evaluation approach is to do user studies, asking users what they think about automatically generated playlists. Pauws and Eggen (2003) produced an automatic music playlist generator called PATS (Personalized Automatic Track Selection) and did a controlled user experiment to compare the quality of PATS-generated playlists with randomly assembled playlists. The user experiment included a short questionnaire and a post-experiment interview. PATS playlists beat randomly assembled playlists, as they contained more preferred songs and had higher

ratings. In order to evaluate music recommender systems and determine the factors that influence evaluations, Barrington, Oda, and Lanckriet (2009) built a new platform for people to evaluate music recommendations. However, user studies are time consuming and expensive, so it is hard to conduct studies of sufficient size to achieve statistically meaningful results. For example, in Pauws and Eggen's study, there were only 20 participants. Barrington and fellow researchers recruited 185 subjects to take part in their experiment. But another problem is that such studies are difficult to reproduce.

Objective Evaluation

One commonly used objective evaluation is to measure the homogeneity or diversity of a playlist. To measure the homogeneity of a playlist, the genre labels can be used to indicate the music similarity (Flexer, Schnitzer, Gasser, & Widmer, 2008). Knees, Pohle, Schedl, & Widmer (2006) estimated the long-term consistency of a playlist by calculating the Shannon entropy of the genre distribution. Another study suggested that diversity is an important quality criteria for playlists (Slaney & White, 2006).

Another way to evaluate objectively is to compare the generated playlist with reference playlists. Reference playlists can be existing playlists extracted from music services or hand-crafted playlists created by music enthusiasts. Some researchers using this approach treat playlist evaluation as an information retrieval problem (Platt, Burges, Swenson, Weare, & Zheng, 2002; Maillet, Eck, Desjardins, & Lamere, 2009), while McFee and Lanckriet (2011) argued that playlist generation can be viewed as a language modeling problem, where songs constitute the vocabulary and playlists are the sentences.

Methods

This research analyzes users' listening histories to find out whether background knowledge about music can predict implicit user preferences. Track co-occurrence in each user's listening history is assumed to reflect user preferences. Audio signal analysis, track metadata, social web data and usage data are used as sources of background knowledge about music.

3.1 Data Collection

Data was collected from Last.fm and Gracenote. Last.fm is a music website. It is well-known for its detailed profile of each user. The information is used for music recommendations. Listening history data, social web data, and usage data were collected from Last.fm. Gracenote is an entertainment data and technology company. It provides music metadata and music recognition technologies. Audio signal analysis data and track metadata were collected from Gracenote. Last.fm was chosen as a source of listening history data, because it records users' listening logs on both the system itself and a wide range of other third-party music and media players (Vigliensoni & Fujinaga, 2014). Gracenote provided audio signal analysis data (not available on Last.fm) and rich metadata for tracks. Last.fm gives access to the music data resources through their public API (Last.fm Web Services, 2018). The Gracenote Web API delivers a rich set of music metadata to help power interactive experiences for any connected application (Web API, 2018). There are also unofficial wrappers for the Gracenote Web API for various

languages, including Python. The Python wrapper abstracts the XML protocol and allow text-based lookups of track metadata, which made the data collection work much easier.

Listening History Data

Collecting listening history data through the Last.fm API requires knowing listeners' usernames in advance. An open data file of Last.fm user profiles (Two Million LastFM User Profiles, 2013) was used as a source of usernames. I looked for listeners with at least 200 tracks listened to between October 2016 and October 2017. The first 10,000 valid users were selected to do the analysis.

I used the `user.getTopTracks` Last.fm API method to get the 200 tracks most listened to by each user. In the response, tracks are ranked based on the number of times they occurred in the user's listening history. With this top 200 tracks data, track co-occurrence can be determined.

Background Knowledge Data

Background knowledge can be classified into categories: audio signal analysis, track metadata, social web data and usage data (Bonnin & Jannach, 2015).

The Python wrapper to the Gracenote Web API takes an artist name and track name as input and returns track information as output. Mood and tempo are extracted from the audio signal. The artist name, artist era (years active), artist origin (city, country, and region), artist gender, album name, album release year, and genre are provided as track metadata.

For the last two categories of background knowledge, I used the `track.getInfo` Last.fm API method to get the top tags (top-5 tags and all available top tags) as social web data, and listener counts and play counts as usage data that can be used to as a proxy for track popularity.

3.2 Data Analysis

Data Preprocessing

Having the top 200 track data for 10,000 users, and the methods available to retrieve background knowledge data for each track, several steps still needed to be done to get track co-occurrence counts and background information for each pair of tracks.

The first step was to transform the dataset from a user-oriented organization into a track-oriented organization. I wrote Python code to go through all 10,000 users to create a list of all unique tracks. The number of times each track occurred and in which users' listening history it occurred were recorded. 775,262 unique tracks were listened to by the sample users. To reduce the number of tracks and make the analysis more efficient, I filtered out tracks that were listened to less than 100 times, assuming that tracks not occurring frequently in users' listening history will not have high co-occurrence with other tracks either. This reduced the number of unique tracks to 830. After having filtered the track list, I used the Gracenote Web API and the Last.fm API to get background knowledge data.

The second step was to analyze track pairs, count co-occurrences, and calculate similarity values using the background knowledge. 830 tracks can make 344,035 track pairs. Track pairs co-occurring less than 10 times were filtered, which resulted in to 42,887 pairs. For

different types of background knowledge, I defined similarity in three different ways. For single string values, for example, artist name and album title, if tracks have the same string value the similarity score is 1, otherwise it is 0. For multiple string values, for example, top tags, genre, and mood, the similarity score is the number of common string values. For example, if the genre for track 1 is “Pop”, “Pop Vocal”, and “Western Pop”, and for track 2 it is “Pop”, “Dance Pop”, and “Western Pop”, the common string values for track 1 and track 2 are “Pop” and “Western Pop”, so the similarity score is 2. For numeric values, for example, album year, listeners, and playcount, I calculate the similarity score as the difference of two tracks. But we need to keep in mind that lower difference value reflects higher similarity score. For example, the difference of album year for track pair 1 is 3 years, while the difference of album year for track pair 2 is 1 year. Track pair 2 has the higher similarity score. For album year, listeners and playcount, I also calculate the harmonic mean of two tracks to represent the average value of album release year and popularity. The harmonic mean formula is shown as follows.

$$H = \frac{2 \times X_1 \times X_2}{X_1 + X_2}$$

The reason why harmonic mean is chosen here instead of arithmetic mean is that harmonic mean is more sensitive to small values. For example, if one track has few listeners, while the other track has great number of listeners, I do not want their pair popularity to be overestimated.

The third step is to do the normalization. Social web data and usage data need this extra step. For social web data, I get both top-5 tags and all-available-top-tags to see which one works better in the correlation analysis. However, the all-available-top-tags needs to be

normalized, because the number of tags for each track varies a lot. Track pairs with more tags have a higher chance to have more common tags, even when those two tracks are not similar. To address this, I divide the number of all-available-top-tags by the average length of tags for the track pair to get toptags_norm. The average length is calculated by harmonic mean. For usage data, the number of listeners and play count for the tracks are counted over years. It means that, assuming two tracks with similar popularity, a track released 10 years ago tends to have many more listeners and a higher play count than a track released last year.

	listeners	playcount
album_year	-0.811	-0.723

Table 1. Correlation between album release year and usage data.

Table 1 shows that the correlations between album release year and listeners and between album release year and playcount are -0.811 and -0.723 respectively, which is very strong. So, I divide the number of listeners and play count by the number of years from the album released (2018 minus the album release year) to normalize the values.

Correlation and Regression Analysis

My research questions are restated here:

Is there relationship between background knowledge about music and track co-occurrence frequency?

What type of background knowledge about music best predicts track co-occurrence frequency?

In order to answer the research questions, I did correlation analysis between background knowledge data and track co-occurrence frequency and regression analysis to predict the

track pair co-occurrence. The independent variables are the similarity scores for track pairs, calculated based on background knowledge. The dependent variable is the co-occurrence frequency of the track pair. Since all variables are transformed into ratios, correlation is the appropriate method. First, I determined if the correlations were significant. Then, Pearson's r (the Pearson product-moment coefficient) was compared to see which independent variables had stronger correlations. Finally, a multiple linear regression model was built to predict the track pair co-occurrence count.

Background knowledge data was classified into four types: the audio signal analysis, track metadata, social web data and usage data (see the section entitled Data Collection for details). I expected to find significant correlations between similarity scores based on background knowledge and track pair co-occurrence counts. Correlations were expected to be positive for similarity scores. If similarity scores are reflected by difference of two tracks, like album year, listeners, and playcount, correlations were expected to be negative. The strength of correlation for scores based on different types of background knowledge was expected to vary, with some scores having a stronger correlation with co-occurrence count and thus potentially more useful for generating playlists. I also expected to find some correlations between the scores themselves, both those based on the same type background knowledge and those that were not.

Correlation is not enough to find out which variables have a larger impact on the dependent variable. Linear regression is a very powerful data analysis technique. It is a linear approach for modelling the relationship between dependent variable and one or more independent variables. When there is one independent variable, it is called simple linear regression, and the formula is for a line. When there are more than one independent

variables, it is called multiple linear regression. Since the independent variables of this study are more than one, I did multiple linear regression. Linear regression allows us to see which independent variables have a statistically significant impact on the dependent variable, compare the impact of each independent variable on the dependent variable, and predict the dependent variable value when independent values are known. The R square value of the linear regression model tells us how much variation in the dependent variable is explained by all of entered independent variables. However, if there is high correlation between two independent variables, it could cause multicollinearity problem in the multiple linear regression model (Stepwise Regression in SPSS – Example, 2018). When highly intercorrelated variables are entered in the model, the coefficient for predicting dependent variable can be not statistically significant. To resolve multicollinearity, I used stepwise regression. The stepwise method starts with zero predictors (independent variables) in the model, and then adds the strongest predictor if its coefficient for predicting dependent variable in statistically significant ($p<0.05$). Independent variables will be entered one by one based on their coefficients. During the entering process, some previously entered predictors may become not significant, then they will be removed. The process ends when none of the excluded predictors is significant.

Results

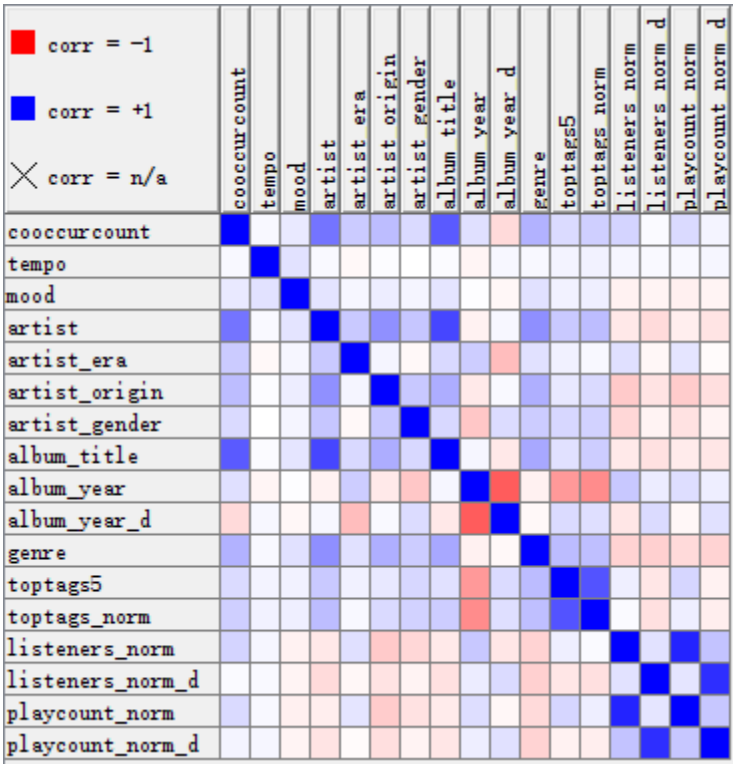

Figure 1. The correlation result for all the variables.

The correlations between co-occur counts and other variables were all significant at the 0.01 level. Album title had the strongest correlation of 0.648 and the difference of listeners has the weakest correlation of 0.02.

Figure 1 shows the overall result for the correlation analysis in a color labeled graph. Blue refers to positive correlation and red refers to negative correlation. The darker the color, the stronger the correlation.

4.1 Correlation

The Audio Signal

Mood and tempo were extracted as audio signal attributes. Example mood descriptions are “Excited”, “Sensual”, and “Intimate”. Example tempo values for a slow song are “Slow Tempo”, “40s”, and “Slow”, while a fast song could have values like “Fast Tempo”, “180s”, and “Very Fast”.

The correlations between co-occurrence counts and mood and between co-occurrence counts and tempo are 0.088 and 0.028 respectively, which are both very weak. The correlation between mood and tempo is significant but is also weak (0.115).

The result shows that track pairs with similar mood and similar tempo do not have a greater or lesser chance of occurring together.

Metadata

Rich information about the artist and album is extracted as metadata for the track. Artist information includes the artist name, artist era (years active), artist origin (city, country, and region), and artist gender. Album information includes its name, its release year, and its genre.

	co-occur	artist name	artist era	artist origin	artist gender
artist name	0.545	1	0.21	0.435	0.225
artist era	0.202	0.21	1	0.04	-0.028
artist origin	0.259	0.435	0.04	1	0.217
artist gender	0.145	0.225	-0.028	0.217	1

Table 2. Correlation between co-occurrence and artist information.

The correlations between co-occurrence and artist information are shown in the second column of table 2. Artist name has the strongest correlation of 0.545. Artist era and artist

origin have correlations above 0.20 (0.202 and 0.259 respectively). Artist gender has the weakest correlation of 0.145, which is below 0.20. So, tracks from the same artist are more likely to occur in the same user's listening history. Tracks from artists that were active during nearby eras or came from nearby regions are also likely to occur together, but the relationship is weaker. However, the artist gender does not contribute much to co-occurrence. Users do not listen to a pair of songs more because the singers are both male or both female.

Among the correlations between different aspects of the artist information, we can observe that because artist name determines the artist era, origin, and gender, the third column for artist name does not give much useful information. In the last three columns, only artist origin and artist gender have a correlation above 0.20. All other correlations are all below 0.20. It seems that there is slight relationship between artist origin and artist gender, while the artist's era is independent from their origin and gender.

	co-occur	album title	album year	album year_d	genre
album title	0.648	1	0.035	-0.09	0.341
album year	0.121	0.035	1	-0.641	-0.056
album year_d	-0.144	-0.09	-0.641	1	-0.029
genre	0.301	0.341	-0.056	-0.029	1

Table 3. Correlation between co-occurrence and album information. Album year_d refers to the difference of track pair's album release year.

The correlations between co-occurrence and the album information are shown in the second column of table 3. Album title has the strongest correlation of 0.648, which is stronger than artist name. The genre of the album has a correlation of 0.301, which is above 0.20. Since album year is a numeric value, the similarity score was calculated as the difference of the tracks' two album years. The harmonic mean (shown as album year in table 3) was also calculated as the average album release year. As I expected, the

correlation between co-occurrence and the difference of the album year is negative. However, the absolute value of the correlation below 0.20. The correlation between co-occurrence and average album release year is 0.121, which is also below 0.20. The result suggests that tracks from the same album are very likely to occur in the same user's listening history. The genre of the album also has some positive correlation. However, there is no suggestion of a relationship between the co-occurrence and the release year of the album.

For the correlations between different aspects of album information, I would like to ignore the column of album title, because album title determines all other album information. The correlation between average album year and genre and between the difference of album year and genre is -0.056 and -0.029. The absolute values are both below 0.10. So there is hardly any relationship between them.

There seem to be some relationships between the artist and album information. Since the album title determines all other attributes, correlations with the album title are not considered here. Among the other attributes, the correlation between the difference of album years and the artist era is -0.257. It shows that tracks released in similar years are likely from artists that are active in similar years, which is what we would expect. The correlation of average album year and artist gender is -0.22, the absolute value of which is above 0.20. It shows that track pair released in recent years are more likely from different gender artists than that released years ago. And the correlation between genre of album and artist name and between genre of album and artist origin are 0.438 and 0.316 respectively, which are both above 0.20. The correlation between genre and artist gender is below 0.20 but is very close to it (0.199). So, tracks in similar genres are likely by the

same artist or by artists from nearby regions. There may be a weak correlation between genre and artist gender as well.

Social Web Data

Top tags data was extracted as social web data. There are two tag-based similarity scores in the study. One is top-5 tags, which is based on the 5 most popular tags assigned by users to the track. The other one is based on all the available tags for the tracks and normalizes the count of common tags by the average number of tags of the track pair.

The correlations between co-occurrence and these two tag-based similarity scores are 0.142 (top-5 tags) and 0.188 (all-available-top-tags). The correlation for all-available-top-tags is slightly stronger, but they are both below 0.20. Tracks having common top tags assigned by users are not likely to occur together more. There is a strong correlation (0.673) between top-5 tags and all-available-top-tags.

Usage Data

Listeners and playcount data were extracted as usage data reflecting track popularity. Listeners and playcount data are both numeric values, so scores were calculated as the differences in values (listeners_d, playcount_d) between the two tracks. However, the actual level of track popularity may be lost with this scoring strategy, so the harmonic mean of the two tracks' values was calculated as an alternative. Both the difference score and the harmonic mean were calculated based on normalized data.

	co-occur	listeners	listeners_d	playcount	playcount_d
listeners	0.167	1	0.114	0.864	0.234
listeners_d	0.02	0.114	1	0.103	0.82
playcount	0.145	0.864	0.103	1	0.221
playcount_d	0.043	0.234	0.82	0.221	1

Table 4. Correlation between co-occurrence and track popularity data.

All correlations between co-occurrence and popularity data are shown in the second column of table 4. The absolute correlation values are all below 0.20, especially with the difference scores which are close to 0. Also, the difference in listener counts and the difference in play counts do not have the negative correlation that I expected.

For the correlations between track popularity data, the correlations between the average listener count and average play count and between the count differences are very strong (are 0.864 and 0.82 respectively). So the number of listeners and the number of plays are correlated, which is to be expected.

Correlation Across Types

For correlations across types, we can look back to figure 1.

Between metadata and social web data, the correlations between artist name and the two tag-based similarity scores are both above 0.20 (0.213 for top-5 tags and 0.258 for all-available-top-tags). The correlations between genre and the two tag-based similarity scores are above 0.25 (0.264 for top-5 tags and 0.252 for all-available-top-tags). The correlations between average album year and the two tag-based similarity scores are negative. The absolute values are above 0.4. So, tracks from the same artist or in similar genres are likely to have similar top tags. Old tracks tend to have more common tags.

Between track metadata and usage data, the correlations between artist origin and average usage data are negative and the absolute values are above 0.20. The correlation between average album year and average number of listeners is 0.214, which is above 0.20. Therefore, tracks from artists coming from different areas tend to have higher average usage data. Tracks from recent years are likely to have more listeners.

4.2 Multiple Linear Regression

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.648[a]	.420	.419	19.782
2	.685[b]	.470	.470	18.908
3	.696[c]	.484	.484	18.645
4	.702[d]	.493	.493	18.489
5	.708[e]	.501	.501	18.342
6	.713[f]	.509	.509	18.193
7	.716[g]	.513	.513	18.120
8	.718[h]	.515	.515	18.075
9	.719[i]	.517	.517	18.051
10	.720[j]	.518	.518	18.029
11	.720[k]	.519	.519	18.010
12	.721[l]	.519	.519	18.005
13	.721[m]	.519	.519	18.003

Figure 2. Model Summary for stepwise multiple linear regression. 13 variables are entered in the order: album title, listeners, genre, the difference of listeners, artist name, the difference of album year, artist gender, artist origin, artist era, toptags(norm), album year, mood, playcount

Using stepwise linear regression training, a model using 13 variables was developed (see figure 2). Although mood and play count entered the model, they hardly improved the R square value, so the final multiple linear regression model was built on album title, listeners, genre, the difference of listeners, artist name, the difference of album year, artist gender, artist origin, artist era, toptags(norm), and album year. The value of adjusted R square is 0.519. Here is the formula of the model.

$$Co-occurrence = 75.32 \times album_{title} + 0.0002 \times listeners + 2.07 \times genre + 6.44 \times 10^{-5} \times listeners_{d} + 11.83 \times artist_{name} - 0.32 \times album_{year_{d}} + 1.70 \times artist_{gender} + 1.11 \times artist_{origin} + 1.87 \times artist_{era} + 11.29 \times toptags_{norm} + 0.22 \times album_{year} - 438.74$$

Tempo, top-5 tags, the difference of play counts, play count, and mood were excluded from the model.

4.3 Evaluation

A simple objective evaluation was done. I compared the predicted track with tracks in the users' listening history.

User 1			User 2			User 3		
No.	Artist	Track Name	No.	Artist	Track Name	No.	Artist	Track Name
1	Lacuna Coil	Blood, Tears, Dust	6	Emily Haines	Fatal Gift	11	Dirty Radio	Champagne Bubbles
2	Led Zeppelin	Custard Pie	7	Goldfrapp	Moon In Your Mouth	12	OFENBACH	Katchi (Ofenbach vs. Nick Waterhouse)
3	Led Zeppelin	Ten Years Gone	8	Goldfrapp	Ocean	13	Britney Spears	Work B**ch
4	Eagles	Wasted Time	9	Austra	I Love You More Than You Love Yourself	14	MOONZZ	Satisfy
5	Lacuna Coil	Downfall	10	Austra	I'm a Monster	15	Katy Perry	Swish Swish

Table 5. Top 5 tracks in 3 users' listening history.

Three valid users, who are not included in the 10,000 training users, were found in the open data file of Last.fm user profiles (Two Million LastFM User Profiles, 2013) to extract test data. They listened at least 200 tracks in 2017. I selected the top 5 tracks in each of the users listening history. So 15 tracks are divided into 3 group based on the listening history (see Table 5). Considering the whole 15 tracks as the pool of tracks, I used the multiple linear model to predict which track would be most likely to occur together with each track (the seed track). Tracks are ranked based on the predicted co-occurrence count. The larger the count, the higher the probability of co-occurring in the same listening history. Tracks having the predicted co-occurrence count less than or

equal to 0 will not be in the rank. If the predicted track and the seed track is in the same user's listening history, it is predicted correctly.

I introduced four measures of accuracy to evaluate the prediction result. The first two measures are whether the track with the highest rank (P@1) and whether one of the top 3 rank tracks (P@3) are correct. The P@3 can be not applicable when only one track have the predicted co-occurrence greater than 0. The other two measures are precision and recall of all the tracks having the predicted co-occurrence greater than 0. Precision measures that from all the predicted tracks, how many of them is correct. Recall measures that from all the correct tracks, how many of them is predicted.

Track	Predicted Track	P@1>0	P@3>0	P(co-occur>0)	R(co-occur>0)
1	5	Y	N/A	1/9	1/4
2	3	Y	N/A	1/1	1/4
3	2	Y	N/A	1/1	1/4
4	15	N	N/A	0/1	0/4
5	1	Y	N/A	1/9	1/4

Table 6. Evaluation result for user 1.

Table 6 shows that using those 5 tracks as the seed track, only one track has the predicted co-occurrence greater than 0. 4 out of 5 those predicted tracks are correct. The average precision is 0.44, and the average recall is 0.2.

Track	Predicted Track	P@1>0	P@3>0	P(count>0)	R(count>0)
6	15, 9, 10	N	Y	4/10	4/4
7	8, 10, 9	Y	Y	4/10	4/4
8	7, 9, 10	Y	Y	4/10	4/4
9	10, 15, 8	Y	Y	4/10	4/4
10	9, 15, 7	Y	Y	4/10	4/4

Table 7. Evaluation result for user 2.

The evaluation result for tracks listened by user 2 is in table 7. 4 out of 5 highest rank predicted tracks are correct, which is the same as the result of user 1. All 5 tracks find at least one correct track in top 3 rank predicted tracks. The average precision is 0.4, and the average recall is 1.

Track	Predicted Track	P@1>0	P@3>0	P(count>0)	R(count>0)
11	15, 14, 9	Y	Y	3/10	3/4
12	15	Y	N/A	1/1	1/4
13	15, 9, 10	Y	Y	3/10	3/4
14	15, 9, 13	Y	Y	3/8	3/4
15	13, 9, 10	Y	Y	4/12	4/4

Table 8. Evaluation result for user 3.

Table 8 shows the result for user 3. All highest rank predicted tracks are correct. The average precision is 0.46, and the average recall is 0.7.

Discussion

Overall, only album title, artist name, artist era, artist origin, and genre have a correlation with track co-occurrence over 0.20. All these variables are in the track metadata category. However, when building the multiple linear regression model, I found that social web data and usage data also contributed a lot to co-occurrence. There also seem to be some relationships between aspects of background knowledge about music. In this section, I discuss the potential reasons for these relationships and the result of the evaluation.

The results for audio signal analysis are surprising. The literature suggests that playlist generation algorithms based on audio-based similarity work well. However, in our experiment, both mood and tempo were uncorrelated with track co-occurrence, so it seems that these factors do not determine the listener's choice of whether to listen to a track. They do not always listen to fast songs (or slow songs). As for mood, it may be that the description vocabulary is too large, which makes it unlikely for track pairs to have common mood descriptions. Among 42,887 track pairs, only 5075 of them have common mood descriptions. There may be other audio signal analysis data that better predict of listeners' preferences, but limited in time and resources, I did not investigate audio signal analysis data other than tempo and mood.

As expected, most of the track metadata have strong correlations with the track co-occurrence. Only artist gender and album years variables (average and difference) are below 0.20, but still above 0.10. As for the variables having the highest correlation values,

album title and artist name, it is commonly known that songs from the same album and songs by the same artist are more likely to be listened to by the same person. People may listen to both of the songs because they have bought the album or because they are fans of the artist. The relatively high correlation of genre with co-occurrence is also not surprising. As for artist origin, it is understandable that listeners tend to listen to songs from artists coming from nearby regions. It is interesting to find that the artist era has a stronger correlation than the difference in album release years. However, the difference in album release years actually enters earlier than artist era when building the linear regression model, which will be discussed later. It seems that listening habits are correlated with the active years of the artist. Also, it is good to see that artist gender does not seem to be correlated with listening habits.

I would like to discuss the relationships between track metadata variables as well. Artist origin has a slight correlation with artist gender. The reason may be that some of the regions tend to have more male artist (or female artist). It is common knowledge that artists who are active in similar years are likely to release album in similar years. The negative correlation between artist gender and the average album release year shows that in recent years people care less about artist gender than old times. And all the artist metadata values except the artist era are related to genre. It seems to be the case that artists have a primary genre. It is true that artists will not necessarily release songs that all belong to the same genre. But generally, most of an artist's songs will be in the same genre. Where the artist comes from and their gender are also correlated with their main genre.

For social web data, I examined top-5 tags and all-available-top-tags (with normalization by average tag length). By comparing those two variables, I hoped to see if the top 5 tags for a track are enough to describe it or if more tag information is better. The results showed that the latter is more likely to be true. However, neither has a strong correlation with track co-occurrence, which means the tags created by listeners for tracks do not seem to be related to listening habits. Since all-available-top-tags includes all the information of top-5 tags, they are closely related to each other.

The average usage data (average listener count and average play count) have the same level of correlation as top tags variables, which are between 0.14-0.20. The differences of usage data (the difference in listener counts and the difference in play counts) for track pairs have low correlations with track co-occurence and are not negative as I expected. Popularity data is mentioned in the literature frequently, but the correlations were not as strong as I expected. Yet they entered the linear regression model relatively early, which will be discussed later in this section.

There are really strong correlations between listener and play counts and between the difference in listener counts and the difference in play counts. Listener count refers to the number of users who have listened the track. Each user can only be counted once. Play count refers to the number of times the track has been listened to. Each user can contribute multiple times to the play count. Thus the listener count reflects the how widely the track was listened to, while the play count variable reflects both how widely and how frequently the track was listened to. That is why those two counts (no matter whether treated as an average or a difference) are closely related. The difference in play

counts also has a slight correlation with the average listeners (0.234) and average play count (0.221), while the difference in listeners does not.

Relationships among background knowledge categories can be divided into two parts: relationships between track metadata and social web data, and relationships between track metadata and usage data.

Among the track metadata, average album year, artist name, and genre have relatively strong correlations with the tag-based similarity scores.

Examples	Top Tags
1	"pop" (genre), "british", "fallon", "airplane", "uk number one", "The X Factor", "talent show", "2016 single" (album year), "bbc radio1 playlist 2016"
2	"pop", "2016" , "future bass" (genre), "the chainsmokers" (artist name), "phoebe ryan"
3	"punk rock", "punk" (genre), "rock", "pop punk" (genre), "Blink 182" (artist name), "alternative" (genre), "90s", "all the small things", "alternative rock", "pop", "american"...(59 tags in total)

Table 5. Example top tags for tracks.

Table 5 shows some examples of the top tags for tracks. Track 1 and 2 were released after 2010, and track 3 was released in 1990s. Track 3 has much more tags than other two tracks. Tags are labeled if they are related to artist name, or genre. Genre tags occur in all three examples, and artist name tags occurs twice. These examples also suggest an explanation for the finding that for average album year and artist, all-available-top-tags is more strongly correlated than top-5 tags, while for genre, top-5 tags has the stronger correlation. If only top-5 tags are considered, it cannot be found out that older tracks have more tags than recent tracks. In table 5, genre-related tags always occur in the first five tags, sometimes even more than once.

Artist origin has a slight negative correlation with average usage data, which means that two tracks by artists coming from different areas are more likely to have higher average popularity. It is hard to explain, and may just be a chance occurrence. As for the relationship between album year and average number of listeners, this could be due to the way the listener count was normalized. As mentioned in the section on data preprocessing, listener count and play count were normalized by dividing the raw counts by the number of years since the album was released, because of the strong correlation between the album release year and the raw counts. Although this simple normalization method works well, usage data actually do not increase linearly for each year, especially the listener count. So, after normalization, there is still slight correlation between album year and average number of listeners.

In the final multiple linear regression model, all variables having correlations with the track co-occurrence of more than 0.10 were included in the model, except for those strongly correlated with another included variable. Since toptags (norm) was included in the model, top-5 tags was excluded. Similarly, since the listener count and the difference in listener counts were included in the model, the play count and the difference in play counts were excluded. The order in which the variables entered the model was not the same as the rank of the correlation value with the track co-occurrence. Listener count and the difference in listener counts entered surprisingly early in the model (2nd and the 4th respectively). Their correlation ranks among the 11 entered variables were 7th and the 11th. Genre, artist gender, and the difference in album years also entered relatively early compared to their correlation ranks. The reason may be that numeric variables (album year, listener count, and play count variables) work better for building the linear

regression model: their averages or differences vary more than the number of common string values for string list variables.

The result of the simple objective evaluation is surprisingly good. Among 15 test tracks, only 2 of them do not find the highest ranked predicted track in the same listening history. It indicates that our prediction model does work in predicting co-occurred track pairs. The average precision of 3 users is stable at around 0.4, while the average recall varies from 0.2 for user 1 to higher than 0.7 for user 2 and 3. Due to limited time, the test data size is small. For more reliable result, similar evaluation with larger test dataset or more complex evaluation is encouraged to be done.

Conclusion

Correlation and regression analysis was conducted to investigate the relationships between background knowledge about and track co-occurrence frequency and predict the track co-occurrence frequency. The correlations were all significant at the 0.01 level. However, some of the relationships are too weak to enter the multiple linear regression model. Considering both the correlation value and the contribution to the linear regression model, track metadata and usage data have the strongest relationship with track co-occurrence frequency.

The result also shows some relationships between genre and other track metadata and between top tags and track metadata. Artists have their main genre. Furthermore, their main genre is likely to be correlated with where they come from and their gender. Having a close look to each track's top tags, I find that tracks released years ago tend to have more tags created by users than recent released tracks. Also artist name, and genre are frequently mentioned in the top tags, especially genre. Genre related tags always occur in the top-5 tags.

The objective evaluation shows that the final multiple linear regression model works well in predicting track co-occurrence frequency.

There are several limitations of the study.

First, although all four categories of background knowledge are covered, the track information extracted for each category was not complete. For example, beside tempo and mood, audio signal analysis data includes danceability, energy, and speechiness.

As mentioned in the discussion section, the normalization of usage data in the study is very simple, based on the assumption that usage data increases linearly for each year. However, this is not true in reality. More complex normalization could be applied to reach a more accurate result.

Due to time and resource restriction, a simple objective evaluation was done for the multiple linear regression model. Only 15 tracks from 3 users' listening history was extracted as test data. Similar evaluation with larger test dataset or more complex evaluation is encouraged to be done. For more complex evaluation, I suggest a hybrid evaluation method which combines subjective evaluation and objective evaluation.

Bibliography

Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In Acm sigmod record (Vol. 22, No. 2, pp. 207-216). ACM.

Agrawal, R., & Srikant, R. (1995, March). Mining sequential patterns. In Data Engineering, 1995. Proceedings of the Eleventh International Conference on (pp. 3-14). IEEE.

Baccigalupo, C., & Plaza, E. (2006, September). Case-based sequential ordering of songs for playlist recommendation. In European Conference on Case-Based Reasoning (pp. 286-300). Springer, Berlin, Heidelberg.

Barrington, L., Oda, R., & Lanckriet, G. R. (2009, October). Smarter than Genius? Human Evaluation of Music Recommender Systems. In ISMIR (Vol. 9, pp. 357-362).

Bonnin, G., & Jannach, D. (2013, June). A comparison of playlist generation strategies for music recommendation and a new baseline scheme. In Workshops at the twenty-seventh AAAI conference on artificial intelligence.

Bonnin, G., & Jannach, D. (2015). Automated generation of music playlists: Survey and experiments. ACM Computing Surveys (CSUR), 47(2), 26.

Bull, M. (2006). Investigating the culture of mobile listening: From Walkman to iPod. In Consuming music together (pp. 131-149). Springer, Dordrecht.

Chen, S., Moore, J. L., Turnbull, D., & Joachims, T. (2012, August). Playlist prediction via metric embedding. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 714-722). ACM.

Edison, T. A. (1878). Improvement in phonograph or speaking machines. US Patent Number 200521.

Fields, B. (2011). Contextualize your listening: the playlist as recommendation engine (Doctoral dissertation, Goldsmiths College (University of London)).

Flexer, A., Schnitzer, D., Gasser, M., & Widmer, G. (2008). Playlist Generation using Start and End Songs. In ISMIR (Vol. 8, pp. 173-178).

Hariri, N., Mobasher, B., & Burke, R. (2012, September). Context-aware music recommendation based on latent topic sequential patterns. In Proceedings of the sixth ACM conference on Recommender systems (pp. 131-138). ACM.

Hornung, T., Ziegler, C. N., Franz, S., Przyjaciel-Zablocki, M., Schatzle, A., & Lausen, G. (2013, November). Evaluating hybrid music recommender systems. In Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on (Vol. 1, pp. 57-64). IEEE.

Jannach, D., Zanker, M., Ge, M., & Gröning, M. (2012, September). Recommender systems in computer science and information systems–a landscape of research. In International Conference on Electronic Commerce and Web Technologies (pp. 76-87). Springer, Berlin, Heidelberg.

Knees, P., Pohle, T., Schedl, M., & Widmer, G. (2006, October). Combining audio-based similarity with web-based data to accelerate automatic music playlist generation. In

Proceedings of the 8th ACM international workshop on Multimedia information retrieval (pp. 147-154). ACM.

Last.fm Web Services. (2018). Retrieved from Last.fm: https://www.last.fm/api

Leong, T., Howard, S., & Vetere, F. (2008). FEATURE Take a chance on me: using randomness for the design of digital devices. Interactions, 15(3), 16-19.

Maillet, F., Eck, D., Desjardins, G., & Lamere, P. (2009, October). Steerable Playlist Generation by Learning Song Similarity from Radio Station Playlists. In ISMIR (pp. 345-350).

McFee, B., & Lanckriet, G. R. (2011, October). The Natural Language of Playlists. In ISMIR (Vol. 11, pp. 537-542).

Pampalk, E., Pohle, T., & Widmer, G. (2005, September). Dynamic Playlist Generation Based on Skipping Behavior. In ISMIR (Vol. 5, pp. 634-637).

Pauws, S., & Eggen, B. (2003). Realization and user evaluation of an automatic playlist generator. Journal of new music research, 32(2), 179-192.

Platt, J. C., Burges, C. J., Swenson, S., Weare, C., & Zheng, A. (2002). Learning a gaussian process prior for automatically generating music playlists. In Advances in neural information processing systems (pp. 1425-1432).

Slaney, M., & White, W. (2006, October). Measuring playlist diversity for recommendation systems. In Proceedings of the 1st ACM workshop on Audio and music computing multimedia (pp. 77-82). ACM.

Stepwise Regression in SPSS – Example (2018). Retrieved from SPSS Tutorials: https://www.spss-tutorials.com/stepwise-regression-in-spss-example/

Two Million LastFM User Profiles. (2013). Retrieved from Socrata: https://opendata.socrata.com/Business/Two-Million-LastFM-User-Profiles/5vvd-truf

Viglensoni, G., & Fujinaga, I. (2014, July). Identifying time zones in a large dataset of music listening logs. In Proceedings of the first international workshop on Social media retrieval and analysis (pp. 27-32). ACM.

Wall, T. (2007). Finding an alternative: Music programming in US college radio. Radio journal: International studies in broadcast & audio media, 5(1), 35-54.

Web API. (2018). Retrieved from Gracenote Developer Music + Auto APIs: https://developer.gracenote.com/web-api

Weber, W. (2001). From miscellany to homogeneity in concert programming. Poetics, 29(2), 125-134.