Cross-Language Music Recommendation Exploration

Stefanos Stoikos
st.stoikos@gmail.com
Pomona College
Claremont, California, USA

David Kauchak
david.kauchak@pomona.edu
Pomona College
Claremont, California, USA

Alexandra Papoutsaki
alexandra.papoutsaki@pomona.edu
Pomona College
Claremont, California, USA

Douglas Turnbull
dturnbull@ithaca.edu
Ithaca College
Ithaca, New York, USA

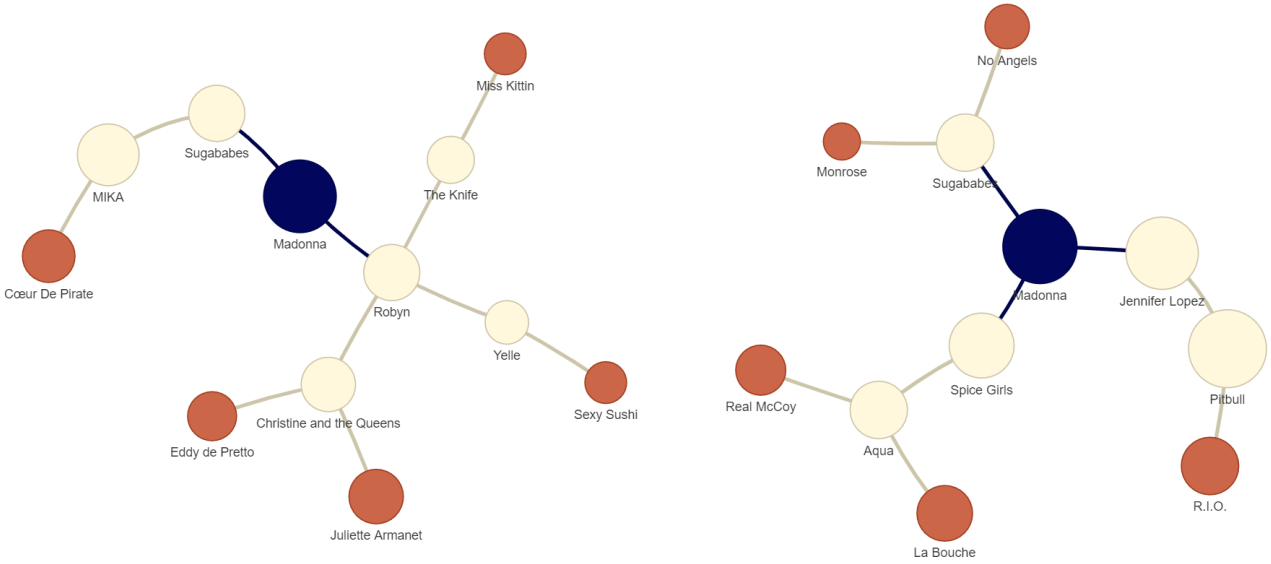

Figure 1: Demo: Breadth-First Search algorithm visualization for searching French and German artist recommendations stemming from Madonna. The left one is for French and the right one is for German. The blue nodes represent the base/root artists, the beige nodes represent the transitive/related artists, and the orange nodes represent the top-5 matching artists in the respective target language.

ABSTRACT

Recommendation systems are essential for music platforms to drive exploration and discovery for users. Little work has been done in exploring cross-language music recommendation systems, which represent another avenue for music exploration. In this paper, we collected and created a database of over 200,000 artists, which includes subsets of artists that sing in 8 different languages other than English. Our goal was to recommend artists in those 8 other language subsets for a given English-speaking artist. Using Spotify's API-related artists feature, we implemented two approaches: a matrix factorization model using alternating least squares and a breadth-first search system. Both systems perform significantly better than a random baseline based on accuracy of the base artist's genre with the breadth-first search model outperforming the matrix factorization technique. We conclude with suggestions for improving the performance and reach of cross-language music recommendation systems.

ICMR '23, June 12–15, 2023, Thessaloniki, Greece

ACM ISBN 979-8-4007-0178-8/23/06.
https://doi.org/10.1145/3591106.3592274

CCS CONCEPTS

• **Information systems** → *Recommender systems*; **Music retrieval**; **Multilingual and cross-lingual retrieval**.

KEYWORDS

graph algorithms, recommendation systems, music recommendation, cross-language retrieval

ACM Reference Format:
Stefanos Stoikos, David Kauchak, Alexandra Papoutsaki, and Douglas Turnbull. 2023. Cross-Language Music Recommendation Exploration. In *International Conference on Multimedia Retrieval (ICMR '23), June 12–15, 2023, Thessaloniki, Greece.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3591106.3592274

1 INTRODUCTION

Millions of people listen to music at any given moment. While listeners have the opportunity to discover new artists and music through streaming services like Spotify, Pandora, and Apple Music, little work has been done in the area of cross-language recommendation systems. Existing music recommendation systems do not have an explicit mechanism to help people pick up new languages and explore new cultures. In this paper, we introduce an initial exploration into this problem using Spotify's related-artist feature. Using this rich network of connections between artists, we build a model using a collaborative filtering algorithm and compare it against a breadth-first search approach. We define an evaluation metric of artist genre overlap to quantify the performance of the models and provide a case study as an example of how the tool can be used.

2 RELATED WORK

To the best of our knowledge, there has not been any previous work that explicitly examines cross-language music recommendations. Music recommendation systems can be separated into a few subcategories [8]:

(1) *content-based:* recommendations that use lyrics and metadata,
(2) *collaborative-filtering:* methods that utilize groups of users' preferences, and
(3) *psychological and emotional:* methods that use mood and other external factors.

For this work we focus on collaborative-filtering methods. These methods were first popularized with the Netflix prize competition [2] and have been used in multiple applications. They rely on the choices of similar users: if user A and user B both rate the same artists similarly, then the assumption is that they have similar music behaviors. Compared to content-based approaches where we compare songs to each other, collaborative filtering compares the behaviors of users.

A range of techniques have been used for collaborative filtering. Memory-based collaborative filtering uses an array of existing user-item ratings to make predictions and compute the probability of an item being liked by one user based on similar users that have already rated this item. This is often done by neighborhood-based models [3] [6]. Model-based collaborative filtering uses deep learning techniques to understand the underlying structures and connections of user-item ratings. The model is then applied to calculate a rating for a user, given their existing ratings in that database [6]. The most common algorithm of model-based approaches is matrix factorization [11].

Despite the existence of different types of recommendation systems it is important to highlight that the medium of music is very difficult to recommend, compared to other domains like clothing products, movies, or hotels [11]. In particular:

(1) Music can be consumed more quickly compared to other forms of media such as TV shows, movies, and books, due to its short length and ease of access.
(2) Music can have a strong emotional effect on listeners [10] making subjectivity a major obstacle in the search for accurate music recommendation models.
(3) Music availability is rapidly expanding: there are nearly 100 million tracks on Spotify [1], and the volume of music that users can consume is limited. Thus the sparsity of user data for given songs can limit the effectiveness of recommendation models [6].
(4) Collaborative-filtering methods tend to be biased towards more popular artists [4]. Therefore, many results lack diversity.
(5) Many current algorithms do not provide an explanation behind their recommendations, which is problematic since the ability to explain a recommendation often aids the user's confidence in that selection [5] [7].

These limitations become even more apparent for cross-language music recommendation systems, since they have to deal with more "distance" between a base artist and a potential matching artist singing in a different language.

3 DATA

3.1 Collecting Data

Using the Spotify API, we started with Spotify's original playlists from their official account. In particular, we collected the "Top 50" playlists from each country, where one of the 8 selected languages is spoken. We then used snowball sampling to collect additional artists using the 'related artists' feature. In addition, in order to create a strong network of artists, we selected Spotify's original playlists for the following genres:

- rap
- pop
- rock
- indie
- r&b
- hip hop

Using this approach, we collected 228,396 unique artists and for each artist, we also stored their related artists, their genres, their popularity, and their top songs. Table 1 shows the number of artists for each of the languages we examined. Some languages like Spanish and German had more artists than others like Greek or Turkish. Each artist had on average 19.52 related artists.

3.2 Data pre-processing

Not all artists were assigned genres from the Spotify API. We utilized Last.fm's API to fill in any missing genres. Since Last.fm is a crowd-sourced platform, we only used tags that were mentioned by at least 15 people to ensure consistency. In addition, for each artist's related artists, we appended themselves to that list. To the best of our knowledge, there is no concrete way to obtain the language that the artist sings from the Spotify API. Therefore, for an artist to be considered part of a language subset, their genres must include the name of that language. For example, an artist with genres like

Language	Artists
German	4481
Spanish	4429
French	3204
Japanese	2919
Italian	2482
Dutch	1739
Turkish	1607
Greek	643

Table 1: Number of artists per language in our dataset.

'german r&b, german hip hop' would be considered for the german subset of the dataset. To be as precise as possible, we also did a manual review of the genres and excluded any genre that was not genre explicitly tied to a language, for example, 'french electronica'.

4 METHODS

We examined two models for cross-language music recommendations and compared them using artist genre overlap accuracy.

4.1 Matrix Factorization Model (ALS)

We created a sparse 228,396 by 228,396 matrix, where an entry was 1 if there was a relation between the artists and 0 otherwise. We considered each artist to be related to themselves, thus the diagonal of the matrix was filled with 1's. We trained the model using the alternating least squares method[1] with 64 latent factors, 0.001 regularization, and an $\alpha = 20$. These parameters were selected using a grid search. We noticed that increasing the latent factors above 64 caused the model to significantly overfit.

4.2 Breadth-First Search Model (BFS)

Using the collected artists, we created a directed graph where each artist was a node and edges represented a related connection to another artist. To keep the model lightweight, we decided to include the genres only in the 8-language target set of artists. To generate the results, we ran a generic breadth-first search and recorded the depth of the finding.

4.3 Evaluation Metric: Artist Genre Overlap

To quantify the performance of the models, we collected a list of artists from Last.fm's top artist page for 6 genres: rap, pop, rock, indie, r&b, and hip hop. Due to the fact that some artists were placed in multiple genres, we decided to treat each artist and genre pair separately. For example, if artist 'X' was placed in hip hop and rap, we would break it up into Artist 'X' in the genre hip hop and Artist 'X' in the genre rap. This resulted in 123 artists in the six genres. For each method, we generated recommendations for the 123 artists and evaluated the top 5, 10, 50, and 100 results. A recommendation was considered correct recommendation if it was in the same genre as the queried artist. We also include a random baseline.

[1] implementation found at https://github.com/benfred/implicit

5 RESULTS

Figure 2 shows the results for the two different methods compared to the random baseline. Both the breadth-first search algorithm and the matrix factorization approach outperform the random baseline, with breadth-first search (BFS) performing better than alternating-least squares (ALS). The results tend to get diluted as we include more results (from top-5 to top-100 results). We cannot be certain of the reason for this trend, however, we could argue that the artists closer to the base artist tend to be more similar, or in other words, have more overlap in their genres. As we try to recommend more artists, the pool of artists gets diluted with non-similar genre artists.

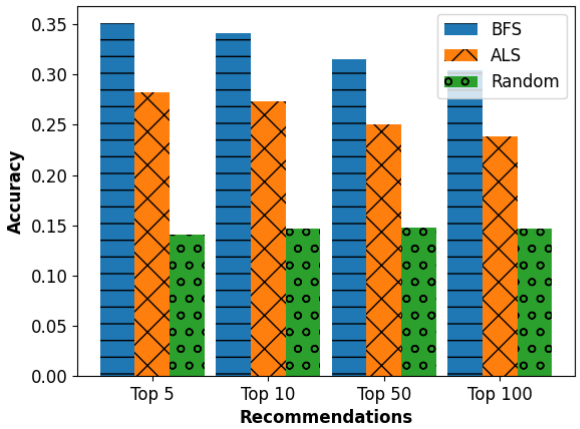

Figure 2: Artist to Artist genre accuracy vs Top-k Recommendations.

Table 2 shows the accuracy by language and genre for the top 5 results for the breadth-first search method. The best recommendations are for Spanish artists with an accuracy of 0.59. With an average depth of 3.71, similar artists are generally 3-4 related artists away from the base artist. Turkish recommendations receive the lowest accuracy of 0.19 with an average depth of 4.75. The best-performing languages tend to find the matching connections before depth 4, whereas the lowest-performing ones tend to have to search further. Overall, the top-performing genre across all the languages is "pop."

Language	Top-5	rap	pop	rock	indie	r&b	hip hop	depth
german	0.45	0.41	0.82	**0.77**	0.52	0.1	0.07	3.34
spanish	**0.59**	**0.7**	0.8	0.52	0.8	0.11	**0.55**	3.77
french	0.4	0.35	0.69	0.19	**0.92**	0.0	0.24	3.79
japanese	0.34	0.33	0.64	0.27	0.28	0.12	0.44	3.99
italian	0.19	0.0	0.67	0.27	0.23	0.0	0.0	4.02
dutch	0.45	0.15	**0.89**	0.72	0.52	**0.36**	0.07	3.45
turkish	0.19	0.38	0.05	0.24	0.04	0.06	0.34	4.75
greek	0.19	0.17	0.3	0.31	0.36	0.0	0.0	4.59

Table 2: Breadth-first search (BFS) results for each language and genre accompanied by the average depth of the artist find

6 DEMO

To further investigate and visualize the performance of the BFS model, we created a live demo [2]. Due to the memory constraints of the web application, the demo currently only contains the subset of artists used for our evaluation metric, artists' genre overlap. In the future, we expect to make the full network available. To use the application, the user selects a base artist and a target language. After they confirm their choice, the breadth-first search model finds the nearest matches in the selected target language. It also provides a visual web map of all of the related artists the algorithm took to find those matches. Hovering over the nodes provides information regarding the artist's genre. Figure 1 shows the search for German and French-speaking artists based on Madonna. The size of the node is proportional to the popularity of the artist based on the Spotify API rating (scale from 0 to 100).

7 CASE STUDY: A BRIDGE BETWEEN GREEK AND SPANISH

To further investigate the breadth-first model and the network itself, we test the significance of one artist making an album in a different language or collaborating with a different language artist.

First, we picked the base language to be Greek and the target language to be Spanish. We ran a test for all 643 Greek artists available in Table 1. For each artist, we took the top 5 Spanish recommendations of the model and recorded the average depth/distance to that recommendation. The mean was *4.4*, the median was *4.2* and the standard deviation was *1.04*. Figure 3 shows a histogram of the results for the 643 artists. Both Shapiro-Wilk and a Kolmogorov-Smirnov test indicate that this distribution is not normally distributed.

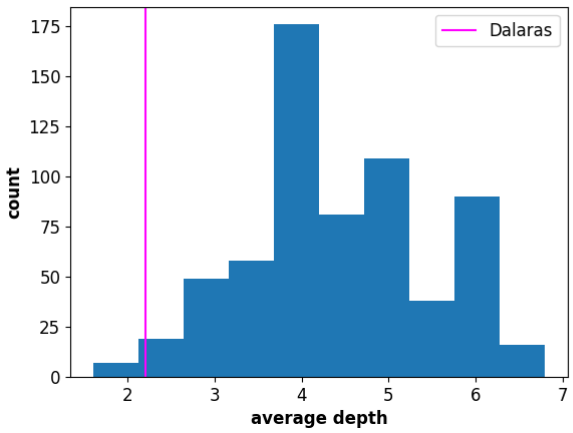

Figure 3: Histogram of average depth from Greek to Spanish Artist; the vertical line indicates the average depth for the Greek artist George Dalaras.

As another example of how the model can be used, we picked the prominent Greek Artist: *George Dalaras*, who besides working on Greek songs, found success in 1987 when he released the album "Latin" that contained his adaptations of some Spanish songs [9]. The average distance/depth from *George Dalaras* to a Spanish recommendation was **2.2**, which has a z-score of *-2.07*. Even though this distribution is not exactly normal, *George Dalaras* is an outlier.

While it is hard to conclude that this artist is "near" the Spanish language due to an album they made, there is strong evidence to suggest that artists that work with multiple languages can create "bridges" between different music cultures that bring listeners closer. We believe that there are more similar types of occurrences that have not been explored yet.

8 DISCUSSION

While matrix factorization techniques have produced very good results in topics ranging from movies, products, and music, we found that a simpler breadth-first search approach had a superior performance on the cross-language music recommendation task. We believe that stems from the fact that the distance between artists of different languages is often too large and that introduces a lot of noise when it comes to training the Matrix Factorization model. This became apparent when we saw that the confidence given by the Matrix Factorization model was less than one percent for the majority of the results. The limited amount of data for languages like Greek or Turkish might have played a role in their relatively poorer performance compared to languages like Spanish or German. However, it could be true that the nature of the related-artist network is such that exemplifies peoples' listening habits. In other words, some languages, like English and Spanish, are near in proximity, in contrast to English and Italian. While the breadth-first search (BFS) model is able to find those nearest connections in 3-4 "hops" away from the root artist, it might be useful to fine-tune each language differently by supplying some sort of weighted edge advantage to specific connections. Nevertheless, as seen with the case study, the breadth-first search highlights the connections that lead to the recommendations, and therefore implicitly explains the results. In addition, we also expect that by including more artists, the network will become more connected and thus give better recommendations.

9 CONCLUSION

In this paper, we introduced the idea of cross-language music recommendations and explored two approaches that leverage artist-relatedness: breadth-first search, and matrix factorization via alternate least squares. We evaluated the models based on genre accuracy of suggested artists and found that both methods outperform a random baseline with the Spanish language performing the best. We provided a live demo where users can test and visualize the breadth-first search model, using the subset of data we used to evaluate its performance. This technique provides clarity to the "explainability" issue that a lot of recommendation systems are facing. Finally, we examined a case study between Greek and Spanish music that points to evidence that artists who work with multiple languages create "bridges" of audiences. There are still many open questions to be explored and we see cross-language music recommendation as an interesting area for future investigation.

[2] https://stefanos-stk-demonstration-app-9ub71w.streamlit.app/

REFERENCES

[1] 2023. About Spotify. https://newsroom.spotify.com/company-info/

[2] James Bennett, Stan Lanning, et al. 2007. The netflix prize. In *Proceedings of KDD cup and workshop*, Vol. 2007. New York, 35.

[3] Robin Burke. 2002. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction* 12, 4 (2002), 331–370.

[4] Oscar Celma and Pedro Cano. 2008. From hits to niches?: Or how popular artists can bias music recommendation and discovery. *Proc. of the 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition* (08 2008). https://doi.org/10.1145/1722149.1722154

[5] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. 241–250.

[6] Ying Qin. 2013. *A historical survey of music recommendation systems: Towards evaluation.* McGill University (Canada).

[7] Rashmi Sinha and Kirsten Medhurst. 2002. The Role of Transparency in Recommender Systems. *Conference on Human Factors in Computing Systems - Proceedings* (05 2002). https://doi.org/10.1145/506443.506619

[8] Yading Song, Simon Dixon, and Marcus Pearce. 2012. A survey of music recommendation systems and future perspectives. In *9th international symposium on computer music modeling and retrieval*, Vol. 4. 395–410.

[9] Philip Sweeney. 1992. *The Virgin Directory of World Music.* Henry Holt.

[10] yi-hsuan Yang and Homer Chen. 2012. Machine Recognition of Music Emotion: A Review. *ACM Transactions on Intelligent Systems and Technology (TIST)* 3 (05 2012). https://doi.org/10.1145/2168752.2168754

[11] Ruisheng Zhang, Qi-dong Liu, Chun-Gui, Jia-Xuan Wei, and Huiyi-Ma. 2014. Collaborative Filtering for Recommender Systems. In *2014 Second International Conference on Advanced Cloud and Big Data*. 301–308. https://doi.org/10.1109/CBD.2014.47