Can We Determine Artist Origin from Past Live Events?

Michael Zhou
Cornell University
mgz27@cornell.edu

Thorsten Joachims
Cornell University
tj@cs.cornell.edu

Douglas Turnbull
Ithaca College
dturnbull@ithaca.edu

ABSTRACT

We explore the task of predicting artist origin (hometown, current city of residence) based on their past live music events, using a heuristic approach; specifically, by calculating the proportion of the artist's events that are located near the city where the artist performs most frequently.

1. INTRODUCTION

Over the past few years, we have been developing a locally-focused music recommendation system called Localify.org [1] . The goal of the system is to help promote music by artists that either originate from or are based out of a specific city that is of interest to the user. While this artist origin information may be available for popular artists via Wikipedia or other sources of biographical information [1], it is often not easy to find for most relatively obscure local artists. In this paper, we explore how knowledge of past live-music events can be used predict artist origin for these less well-known, long-tail artists and use them to mitigate the rich-get-richer bias in music recommendation results [2].

Most previous research on predicting artist origin has focused on web-scraping and natural language processing (NLP) techniques. One particular approach parses origin info via artist's biographies, knowledge graphs, or artist info pages, or search engine methods [1] [3]. However, scraping from artist biographies is less precise since NLP has to be used to extract the origin entity from varying phrases in the biography. While the artist info pages are often available for well-known artists, such information is less common for obscure artists.

In this paper, we explore the feasibility of predicting an artist's origin based on their historical events. Our hypothesis is that when an artist first starts out, they will performs in or near their origin. They will gradually begin to tour first regionally, then nationally, and finally internationally as they becoming increasingly well-known. An artist may also periodically come back for "homecoming" events. Hence, we hypothesize that an artist's earlier events are more significant than their later events, and whichever city the artist most frequently plays in could be a good estimate of their origin.

[1] "https://www.localify.org

2. EXPERIMENTAL SETUP

To use a predicted artist origin in Localify, our Localify system requires a high level of *precision* (that artist is actually from the city that we predict.) Our goal is to get as much *coverage* as possible by making predictions for as many artists as possible while maintaining high precision. To measure this, we use past event histories collect from the BandsInTown (BIT) API [2] to make predictions and compare them against curated ground-truth biographical information from BandInTown or Wikipedia.

BIT started listing past events for artists in early 2013 so many artists that having been playing before then have incomplete event histories. To this end, we only consider artists who have their first event listed after January 1st, 2014. We obtain from our local database a sample of 4766 artists whose first events come after 2013 and are from cities with 10 or more confirmed artists according to BIT or Wikipedia. As some artists may have multiple confirmed origins defined in the database, we only considered artists with one unique confirmed origin. These 4766 artists span from 249 different cities.

Some heuristic indicators we considered were the proportion of events within a certain radius of an artist's most frequent *mode* city, the earliest date when an artist's first event takes place according to BIT, etc. We evaluated the heuristic model the following way: for each of the 4766 artists, we get all events or the first e events of an artist (where $e \in \{10, 20, 30, all\}$; we get the mode city of these events, and then calculate the percentage of events p that are located with a radius r of the mode city. We only committed to a predicted origin label if the artist has at least 5 or 10 events total and p was greater than some threshold t. The results are presented in table Table 1.

We then calculated the *precision*, the number of artists whose predicted label is within the specified range of 10 miles of the ground truth origin divided by the number of artists with a prediction label, and the *coverage*, the number of artists with a prediction label divided by the size of the sample of artists (4766). We repeated this procedure for artists with 10 or more events. Artists with less than 5 events are filtered out due lack of sufficient data.

[2] https://rest.bandsintown.com

3. RESULTS

The precision is mostly between 0.8 and 0.9 for artists having 10 or more events, and always under 0.8 for artists with 5 or more events. Only artists with 10 or more events are shown in Table 1, since there is likely a lack of completeness in event data for artists with just 5 events, as some artists' events might not be recorded in BIT.

To evaluate performance of this heuristic approach, we are interested in having the maximum coverage with at least 0.85 precision. Using this definition, the "best" model consists of artists with at least 80% of their first 20 events in or near their origin city, with a precision of 0.873 and a coverage of 0.025 (118 out of 4766 artists predicted).

We have found 15 total errors in the "best" model. When doing error analysis, we noticed two sources of errors: artists who could be local to multiple origin cities and artists whose earliest events were not found in the BIT data. Out of the 15 errors, 5 of these artists had a strong connection to the predicted city (e.g., relocated from, born in), and 10 of these artists do not have their earliest events recorded in BIT.

Artists who relocated to another city can be cleaned up in the database by adding the relocated origin label to that artist in the database, or simply switching the label to the relocated origin city. An artist who lives and plays frequently somewhere is familiar or "local" to that place, even when he or she relocates. If we do not consider any artist who relocated to the predicted city as incorrect examples, the precision increases to 0.915.

We came up with a way to determine artists whose inaugural events are not recorded in BIT: filtering out artists who have been active before January 1st, 2014 according to Wikipedia, since BIT recorded event data in 2013. Though we thought that applying this filter would improve performance, the precision and coverage both plummeted.

For the other heuristic models, there might be a very small number of examples that may contain mislabeled origins or events. Artists whose data is labeled erroneously in BIT can be filtered out, and those mislabeled in the database can be corrected.

We have also tried using a Naïve Bayes models to predict artist origin. We thought that this model would be effective since it reflects class conditional probabilities (i.e., having many events from a small town is more evidence than having the same number of events from a large city) that an artist is from a specific city. However, this model as well as other supervised learning models (SVMs, Neural Networks) to learn the heuristic features, but none of these methods yielded promising results.

4. CONCLUSION AND DISCUSSION

We proposed a high-precision heuristic approach of predicting an artist's origin based on the artist's historical event data. Specifically, we found the most frequently played city by each artist, and predicted artist origin based on the proportion out of all events or the first e events are above a heuristic threshold t. We found that while the best approach achieves sufficient precision, despite low coverage of artists within the database.

Set of Events e \ Threshold t	0.8	0.85	0.9	0.95
all	0.840	0.842	0.837	0.810
first 30	0.867	0.864	0.86	0.833
first 20	**0.873**	0.884	0.881	0.867
first 10	0.793	0.810	0.810	0.831

Set of Events e \ Threshold t	0.8	0.85	0.9	0.95
all	1.700	1.196	0.902	0.441
first 30	2.056	1.385	1.049	0.504
first 20	**2.476**	1.804	1.238	0.629
first 10	4.050	2.434	2.434	1.238

Table 1. Precision p table (above) and Coverage % table (below) for artists from cities with 10 or more confirmed artists. Each artist has 10 or more events and their first event recorded in BIT came after January 1st, 2014.

Most artists that are correctly predicted are obscure artists who usually stay close to their origin rather than tour around the globe. The BIT API may not contain all the historical past events of many artists, especially for artists that have been active before BIT started recording event data in 2013. The actual first events of most artists may be played in obscure locations such as bars and restaurants, and it may be difficult to record them online. Some techniques to improve this include crowd-sourcing, where anyone who has attended inaugural, obscure events by artists can post event data online. In addition, if an artist's info page is already available on Wikipedia, it is more accurate to directly extract the artist's origin from Wikipedia instead of predicting heuristically.

Our future work includes improving artist origin predictions by combining information extracted from past event histories and artist biographies. By improving artist origin prediction we will be better able to help users connect with local artists and support their local artistic communities.

5. REFERENCES

[1] M. Schedl, K. Seyerlehner, D. Schnitzer, G. Widmer, and C. Schiketanz, "Three web-based heuristics to determine a person's or institution's country of origin," in *ACM SIGIR*, 2010, pp. 801–802.

[2] V. Hurtado, T. Joachims, and D. Turnbull, "Changing how we value local music using personalized recommendation," in *ACM Tapia*, 2020.

[3] S. Govaerts and E. Duval, "A web-based approach to determine the origin of an artist.." in *ISMIR*, 2009, pp. 261–266.