PLACING MUSIC ARTISTS AND SONGS IN TIME USING EDITORIAL METADATA AND WEB MINING TECHNIQUES

Dimitrios Bountouridis, Remco C. Veltkamp, Jan Van Balen
Utrecht University, Department of Information and Computing Sciences
{d.bountouridis,r.c.veltkamp,j.m.h.vanbalen}@uu.nl

ABSTRACT

This paper investigates the novel task of situating music artists and songs in time, thereby adding contextual information that typically correlates with an artist's similarities, collaborations and influences. The proposed method makes use of editorial metadata in conjunction with web mining techniques, aiming to infer an artist's productivity over time and estimate the original year of release of a song. Experimental evaluation over a set of Dutch and American music confirms the practicality and reliability of the proposed methods. As a consequence, large-scale correlational analyses between artist productivity and other musical characteristics (e.g. versatility, eminence) become possible.

1. INTRODUCTION

Many real-world music collections show a lack of metadata when it comes to placing their constituent music entities in a semantic or quantitative context. As a result, content management and disclosure become challenging tasks. Meanwhile, in the emerging field of the digital humanities, well-documented collections are becoming increasingly essential for high quality research.

The lack of contextual information has been typically addressed by content-based approaches, where knowledge is extracted after the actual audio is processed and analysed. Contrarily, Web-Music Information Retrieval (MIR) techniques exploit the "wisdom of the crowd" and use the Web or music metadata hubs in order to estimate the desired information. Typical applications include artist similarity [3, 9], classification [10], country of origin determination [8, 11] and many more exceeding our scope.

This paper focuses on quantifying artists' productivity over time, and estimating the original release time of songs. The productive period of a music artist is important information that is typically highly correlated to his style, influences and similarities to other artists. Teitelbaum et. al. [12] have shown that the activity span is strongly associated with artist collaborations. As such, the productive years constitute a reliable, additional feature for various MIR tasks: as the authors of [13] argue, listeners typically show a certain affection for music related to particular periods of their lives, and therefore time information could act as a basis for music recommendation.

The practical applications of productivity profiles exceed the MIR domain. According to [5], productivity in absolute terms may be the most important factor for a comprehensive understanding of the creativity in music. Based on that, Kozbelt [1] investigated the correlation between productivity and musical characteristics such as versatility and eminence. From a musicological perspective, a quantitative representation of productivity can offer valuable insights in music trends, significant musical events, significant social events, and the mutual influence that may exist between them.

1.1 Problem Definition and Related Work

We define a song's year of release as the year on which it was first released in a recording. We further define as an artist's productivity profile (APP) the distribution of years in which the artist was alive and musically active, meaning recording and releasing albums, singles, etc. The productivity of an artist for a given year corresponds to the number of recorded songs released throughout that year.

Time information regarding an artist's output is typically provided by services and hubs such as MusicBrainz [1], Last.fm [2], Allmusic.com [3] etc. in the form of editorial metadata, i.e. data related to prescriptive knowledge about the music [6], in addition to domain-free sources such as Wikipedia. However, when dealing with old and lesser-known artists, any provided information is highly probable to be erroneous, incomplete or even non-existent. For example, the MusicBrainz profile for the popular Dutch singer Willy Derby (1886-1944) includes a series of compilation albums, released after his death, and only 16 singles out of his huge catalogue. To the best of our knowledge, the only published work aiming at automatically providing time information for artists and songs, although inside a recommendation framework, is [4]. Bogdanov and Herrera address the issue of determining a record's original epoch, meaning the years when the music was first recorded, produced and consumed. This task is handled by

[1] musicbrainz.org

[2] ww.last.fm

[3] www.allmusic.com

finding the release on Discogs[4] with the earliest date and by propagating it using a decreasing weighting scheme. Similarly to MusicBrainz though, Willy Derby's Discogs profile is limited (2 albums, 3 singles and one compilation) and therefore this approach is destined to fail.

The only content-based method we know [13] acts as a proof of the Million Song Dataset's applicability, and aims at estimating a song's year of release based on its audio features. Such an approach shows a small mean error of 6 years but actually estimates the year that the song would best fit in and not its actual year of release. Considering that audio data for old and lesser-known artists, such as Willy Derby, is typically hard to find, this method can be rendered useless in this context. The same holds for audio fingerprinting methods such as [7] and commercial services such as Shazam[5], which in addition face the metadata scarcity problem.

Based on the previous, the need for a reliable method that overcomes the scarcity of the editorial metadata of lesser-known artists and is based on high-level metadata only (e.g. artist name - song title), becomes apparent.

1.2 Contribution

The contribution of this paper is multiple. First, it introduces the novel task of accurately placing music entities, and especially artists, in a time context. In addition, we provide a publicly available testset. Secondly, it employs a methodology that combines both editorial metadata and web mining techniques, a conjunction rarely investigated. The fusion of the different sources is aided by musically meaningful heuristics offering room for research. Thirdly, our method incorporates generic techniques for birth-date and death-date estimation that can find applications outside the music context.

The remainder of this paper is organized as follows. Sections 3 and 4 describe our proposed method for the artist productivity profile and year of release estimation respectively. Sections 5 and 6 summarize the evaluation and experiment results, while section 7 presents our conclusions.

2. GENERAL FRAMEWORK

Our generic method for determining a person's productivity profile (singer, actor, author etc.) is performed in various steps. Web search engines are initially mined to provide the "preliminary" productivity profile. Secondly, editorial metadata hubs are queried for time data about the person's life and works. This generates what we call a "shaping" profile. The information gathered from this process is merged and applied on the first, to attenuate any noise and shape the final productivity distribution (see Figure 1). In the possible case of absent lifespan editorial metadata, our method estimates the birth and death dates based on the preliminary profile.

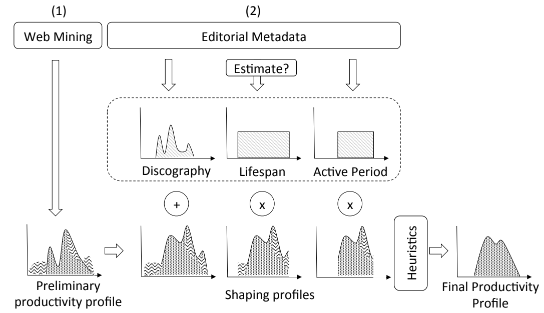

Figure 1. A graphical representation of the artist productivity profile estimation process. The initial Web-mined profile is subsequently noise-filtered, based on editorial (or estimated) medatada and heuristics.

3. DETERMINING ARTIST PRODUCTIVITY PROFILE

3.1 Editorial Metadata Retrieval

Given A an artist name and S^A a set of song titles corresponding to the artist's recordings, we first try to match any of the tuples $\langle A, s_j \rangle$ where $s_j \in S^A$ to the databases of Last.FM, EchoNest and MusicBrainz. Besides being well established and previously used by MIR researches, the employed metadata hubs provide convenient APIs. The following pieces of information are desirable for each artist: *a)* discography, *b)* lifespan and *c)* active years period.

Given that we have retrieved two values corresponding to the start and end years (s, e) of an artist's activity, we create a 130-bin profile $P_{\text{act}} \in [0, 1]^{1\times130}$, spanning the years 1880 to 2010, with both $P_{\text{act}}(s)$ and $P_{\text{act}}(e)$ set to 1. Similarly, we create a profile P_{life} for the lifespan data.

A P_{disc} profile is also created for the discography data, but populating it is more sophisticated. The discography data comprises of album names, song titles accompanied by their release date and "release group" information. Our method assumes that some release groups (e.g. singles) are more reliable than others (e.g. compilations) with regard to the original date determination. Given $D = \{(year_1, weight_1),..., (year_n, weight_n)\}$ the retrieved data from MusicBrainz, with $year_i \in \{1880, .., 2010\}$ and $weight_i \in [0, 1]$, the P_{disc} gets the following values:

$$P_{\text{disc}}(y) = \sum_{\forall i: year_i = y} \frac{weight_i}{N_{\text{release group}_i}} \qquad (1)$$

$N_{\text{release group}}$ is a normalization factor corresponding to the number of recordings belonging to that particular release group.

3.2 Web Mining

The second step is concerned with identifying the Web pages related to the artist under consideration. Our method queries Google and Bing with the scheme "A+music" as in [8], and retrieves the 100 and 80 top-ranked URLs, denoted as sets G and B respectively. Fetching and indexing

[4] www.discogs.com
[5] www.shazam.com

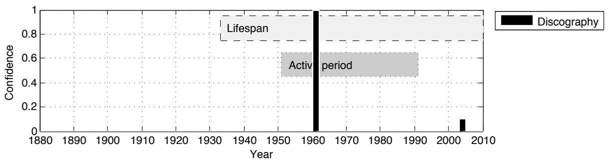

Figure 2. A representation of the three profiles, derived from editorial metadata, for the artist Corry Brokken. Discography data correspond to just one single (1960) and one compilation (2003).

the webpages in $B \cup G$ are performed by the Apache Nutch web crawler [6] and Apache Lucene [7] respectively.

3.2.1 Profile Construction

At this stage we aim at generating a probability distribution that would best model the artist's productivity as it is documented on the Web. This is performed using the following:

- **Returned counts**: The number of returned Lucene pages pc_q for query q, acts a draft estimate of the query's relevance inside the pool of retrieved documents.
- **Co-occurrence analysis:** Considering returned counts as probabilities, the conditional probability of a term t_1 to co-occur in the same page as t_2 can be written as $p(t_1|t_2) = pc_{t_1,t_2}/pc_{t_1}$.
- **Relevance score:** Apache Lucene employs a sophisticated variation of the *tf*idf* and Boolean mode to calculate the similarity between a query q and a document d, denoted $sim(q,d)$. The details of the so-called Lucene Practical Scoring Function [8], are omitted here.

For each year $y_j \in \{1880, 1881, \ldots, 2010\}$ we query Lucene with "$A + y_j$". Incorporating a proximity factor to the query, ensures that when A and y_j are separated by over 300 words, the document containing them would be considered irrelevant. The value of 300 was chosen based on a set of preliminary experiments.

Our method assigns a score to each query "$A + y_j$" using the following formula:

$$s(A, y_j) = \frac{score(A, y_j)}{score(A)} \quad (2)$$

where:

$$score(q) = max_{1<k<pc_q}[sim(q, d_k)] \times pc_q \quad (3)$$

A profile P_{web} of 130 bins is populated such that $P_{\text{web}}(y_j) = s(A, y_j)\ \forall y_j \in Y$. Often, however, time information is not explicitly stated. For instance, it is quite common for an artist that was active during the period 1980-1990, to be considered and identified as an "80's" artist. Based on this, we have identified a set of terms T that semantically correspond to decades. For example, the terms "1960's", "sixties", "60's", "jaren 60" and "jaren zestig" correspond to the period 1960-1970. The two latter country-specific terms are introduced manually but can be automated based on a country of origin estimation process [8,11]. Therefore finally, we query Lucene with "$A + t_j$", where $t_j \in T$, and then increase the value of the ten corresponding bins by $s(A, t_j) \times 0.1$.

3.3 Profile Fusion

By the end of the previous process, the system has acquired four separate profiles (P_{act}, P_{disc} , P_{life} and P_{web}). We combine those pieces of information in a two-step procedure.

It is very likely for the discography to be incomplete for apparent reasons. Our method tries to compensate for that fact by smoothing P_{disc} with a sized-5 Gaussian window. A profile P_{f} is later created such that its bin values hold the weighted sum of the normalized P_{web} and P_{disc} .

The system exploits lifespan and active years data by setting all the coefficients of P_{f} that fall outside the P_{life}, P_{act} boundaries, to zero. Despite this process, the P_{f} data inside the lifespan or active period may still contain a significant amount of noise. Further noise removal is based on the observation that an artist's productivity is usually maximized during his first 20 years of adulthood. This is supported by the dotted line in Figure 3, which represents the distribution of single-type releases across the artist's lifespan, as generated from a Musicbrainz subset of 518, pre-1950's artists. This behaviour is modelled by our method as an envelope-probability density function W (solid line Figure 3). Our envelope's decay slope is less steep in order to accommodate for non-single releases. W is aligned with the artist's birth-date, as provided by P_{life}, and used to weigh P_{f} which constitutes the final artist productivity profile estimate.

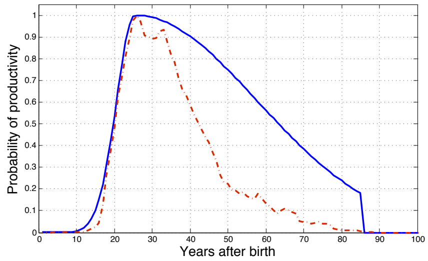

Figure 3. (Solid line) the probability density function W for artist productivity. (Dotted line) the distribution of single-type releases across lifespan.

3.3.1 Birth-Date & Death-Date Estimation

As previously mentioned, it occurs very often that no lifespan information can be found. In this case we use P_{web} which ideally amplifies important years in the life of an

[6] nutch.apache.org
[7] lucene.apache.org/core
[8] lucene.apache.org/core/old_versioned_docs/versions/3_5_0/api/all/org/apache/lucene/search/Similarity.html

artist, represented as sharp peaks. For the birth-date estimate, the idea is to locate those and pick the one that best fits our productivity probability assumptions, as modelled by W. This is achieved by traversing and aligning W with each found peak and then multiplying it with P_{web}. The peak that yields the largest area under the profile is considered the birth-date estimate.

The selection of the peak candidates is a critical procedure, considering that the noise-to-signal ratio can be significantly high. Our method aims at emphasising those peaks that exhibit high surprisingness or unexpectedness, in contrast to noise-generated peaks. This is achieved by scanning the P_{web} profile from left to right; the coefficients of the surprisingness vector $S_v \in [0,1]^{1\times 130}$ are then computed such that:

$$S_v(j) = \frac{P_{\text{web}}(j)}{\sum_{k=1}^{j-1} P_{\text{web}}(k)} \qquad (4)$$

Peak candidates are then selected using a simple thresholding function (Figure 4).

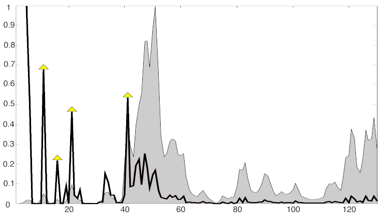

Figure 4. P_{web} and its surprisingness vector S_v (solid line). The more data is processed (from left to right) the less surprising peaks become.

Death-date estimation is based on the assumption that the artist's productivity, as documented on the Web, will be limited or non-existent after his death. Noise however, challenges this assumption; therefore our method uses a peak-picking technique that employs the following pieces of information and heuristics: 1) Birth-date, estimated or known, 2) max life expectancy, set to 90 years and 3) a P'_{web} profile generated with a proximity factor of 10 (instead of 300) which aims at capturing co-occurrences instances of the type "Artist (Year of birth - Year of death)".

The peak picking is done as follows: by using a simple thresholding function on P'_{web}, we firstly pick the peak candidates p. Each one of them is assigned a confidence-probability value $prob(p)$ based on its distance from the birth-date; assuming that peaks far away from the birth-date have higher chance of corresponding to death-date.

Ideally, the peak that maximizes the ratio between distributions to its left and to its right should correspond to the correct death-date. However, considering once again the noise and its non-uniform distribution (e.g. album re-releases after the artist's death), a more sophisticated technique is required. Our method, aiming at capturing the short and long term distributions around each peak, creates four sub-profiles from P_{web}, centred on p, with window sizes of 40 and 10 (P^{40}_{left}, P^{40}_{right}, P^{10}_{left}, P^{10}_{right}). Each candidate p is then assigned a value such that:

$$v_p = \left(\frac{mean(P^{40}_{\text{left}})}{mean(P^{40}_{\text{right}})} + \frac{mean(P^{10}_{\text{left}})}{mean(P^{10}_{\text{right}})} \right) \times prob(p) \qquad (5)$$

The candidate with the maximum v_p is considered the death-date estimate.

We evaluated our lifespan estimation method on a test-set of 100 MusicBrainz artists with known lifespans, ranging from 1880 to 2000. The mean error for the birth-date and death-date was estimated at approximately 11.5 and 10.6 years respectively. This might not be exactly accurate, yet the death-date estimation is not a goal in itself, rather a pragmatic strategy to remove some of the noise in the profiles.

4. DETERMINING YEAR OF RELEASE

Estimating year of release is based on a similar approach as the one employed for artist productivity profiles. The basic idea is to create four separate profiles for each $\langle A, s_j \rangle$: a release information profile from MusicBrainz, a productivity profile for A and two Web profiles for s_j and $A + s_j$ respectively. Processing and fusing them is the final step.

We first query the metadatabases with tuples of the form $\langle A, s_j \rangle$ in order to retrieve the discography. It is now easy to search into the discography of A for any of the songs in S^A. If there is a match, the release group is "Single" and the year of release y is available, then a profile P_{disc} for the song in hand is populated, such that $P_{\text{disc}}(y) = 1$.

During web-mining, two set of queries for each tuple $\langle A, s_j \rangle$ are applied, and more specifically "$s_j + year_k$" and "$s_j + A + year_k$" . Eventually two profiles, $P_{\text{web,s}}$ and $P_{\text{web,s}+A}$, are calculated.

Calculating the year of release estimate is based on the assumption that the input vectors correspond to mixture components. Their conical sum P_{CS} is:

$$P_{\text{CS}} = \mathbf{W_v}^T \times [P_{\text{web,s}}, P_{\text{web,s+A}}, P_{\text{f},A}, P_{\text{disc}}] \qquad (6)$$

with $\mathbf{W_v} = [w_1, w_2, w_3, w_4]$ the weight vector and $P_{f,A}$ the artist's productivity profile. Finding the optimal weights is solved with the employment of a genetic algorithm on a training set. The final year of release estimate is just the P_{CS} coefficient with the maximum corresponding value.

5. EXPERIMENT

5.1 Test Collection

There exists no standardized or previously used data set for this kind of task, therefore we built one from scratch [9] . Note that the requirements are very specific: on the one hand, an evaluation dataset of artist productivity profiles should only include complete or near-complete discographies, and on the other hand, should focus enough on the more challenging artists for which no complete catalogues are readily available on the web. The low commercial distribution and rarity of pre-60's music was ideal for our

[9] Available at www.projects.science.uu.nl/COGITCH

purposes; therefore, we manually gathered from our personal music collection, books [2] and "deep" Web sources, 639 Dutch and American song titles, corresponding to 15 artists (see Table 1), accompanied by original release dates, ranging between the period of 1900 to 1959. Overall, 26832 documents were downloaded and indexed for the year of release and 3391 for artist productivity profile including a set of "noise" webpages (irrelevant to the artists themselves).

The difficulty in assessing the "obscurity" of an artist and the amount of effort required to gather and cross-check his complete discography without using the Web, resulted in a rather small artist productivity profile test set. It should be noted that a direct comparison of our method to others is unfortunately impossible. The work of [4] exploits Discogs which offers limited, or non-existent, data for our particular testset. Regarding year of release estimation, the work of [13] uses purely audio features which are almost impossible to acquire, considering our method's prerequisite for music rarity and oldness.

5.2 Evaluation Measures

For the artist productivity profiles the idea is to examine the overlap between the ground truth APP and estimated APP^* distributions. This is achieved by using precision, recall and their harmonic mean, usually called F-measure. In addition, given the mean values of both profiles' Gaussian fits, denoted m and m^* respectively, we compute $Er = |m - m^*|$ as a measure of the error in terms of time context placement.

Given N, the number of songs in S, YoR_{s_j} the true release date of the song s_j and $YoR^*_{s_j}$ the estimate, "Accuracy" for a year-window of size x is defined by the following formula:

$$Accuracy_x = \frac{\sum_{s_j \in S} f_x(YoR^*_{s_j}, YoR_{s_j})}{N} \quad (7)$$

where

$$f_x(t,e) = \begin{cases} 1 & \text{for} \quad |t-e| \leq x \\ 0 & \text{for} \quad |t-e| > x \end{cases} \quad (8)$$

6. RESULTS AND EVALUATION

The results for the artist productivity profile task are presented in Table 1. Our approach shows low error with regard to the time context placement. It is worth examining certain illustrative cases, starting with "August De Laat" (Figure 5), which presents one of the lowest precisions. The artist is well placed into the time context but our approach assumes a considerable amount of productivity from 1940 to 1954. This misbehaviour relies on the fact that even after De Laat's last recording in 1941, he remained active in non-music areas such as theatre [10] .

In contrast to the previous case, "Bob Scholte" shows both accurate time-context placement and distribution modelling (Figure 6). In the case of "Louis Davids" (1883-1939) presented in Figure 7, lifespan or active years information from the metadabases was unavailable. Our method estimated the correct birth and death dates by performing the peak picking algorithm presented in 3.3.1. Filtering the profile by applying the productivity assumptions, as modelled by W, also attenuated a considerable amount of noise right after the artist's birth.

Artist Name	*Precision%*	*Recall%*	*F%*	*Er*
August De Laat	48.36	83.97	61.38	2.45
Bob Scholte	65.1	90.82	75.84	1.15
Kees Pruis	73.86	80.21	76.9	1.45
Lou Bandy	51.9	97.97	67.85	3.95
Louis Davids	50.2	99.85	66.81	0.11
Willy Derby	90.21	81.93	85.87	1.25
B. Schoepen	52.75	85.11	65.13	6.02
Cole Porter	66.33	88.2	75.72	1.94
Corry Brokken	81.89	88.65	85.14	1.21
Eddy Christiani	85.06	78.22	81.49	0.56
George Gershwin	61.31	83.64	70.76	2.66
Harold Arlen	81.6	68.36	74.39	5.69
Jerome Kern	78.11	63.12	69.82	3.95
Richard Rodgers	46.74	97.13	63.11	7.55
Wim Sonneveld	55.74	95.85	70.49	1.01
Mean	**65.5**	**86.135**	**72.665**	**2.56**

Table 1. Precision, recall and F-measure for the 15 artists in the test set.

Table 2 presents the $Accuracy_x$ for the year-of-release estimation task for windows ranging from 1 to 5. The mean error is 2.91 years. As a general evaluation measure we consider $Accuracy_2$, assuming that this level of detail is appropriate for artists of the era 1900 - 1959. Therefore, for a 2-year window around 81% of the cases are identified as hits; significantly outperforming the random, baseline estimation (mean error = 33.4, $Accuracy_2 = 8.92\%$).

Window Size	1	2	3	4	5
Accuracy	0.66	0.816	0.856	0.888	0.907

Table 2. Accuracy for window size ranging from 1 to 5.

7. CONCLUSIONS

The aim of this report has been to determine an artist's productivity profile and a song's original year of release. Our approach is based on the exploitation of editorial metadata from sources such as MusicBrainz and EchoNest, in addition to Web harvested data. The evaluation demonstrates the strength of the proposed method for year of release estimation; around 81% of the estimates fall within a ±2-year window, with a mean error of 2.91 years.

The results for determining productivity profiles are less impressive though it should be noted that the ground truth generation assumes complete knowledge of the artist's discography, which is not always the case. In fact, in the cases for which we are most certain we have the complete discographies, modelling the productivity distribution is accurate.

With our novel methods, it is possible for the first time to perform large-scale, correlational analysis between productivity and various musical characteristics. Therefore, work such as [1, 5], which is based on manually gathered classical music, cannot only be significantly aided but also expanded to include artists from various eras and of varying popularity.

[10] www.thuisinbrabant.nl/personen/l/laat,-august-de

Given certain enhancements and modifications, our approach can be generalized to accommodate non-music domains. Wikipedia instead of MusicBrainz or EchoNest can be employed for determining an author's productivity profile or the original dates of his publications. Metadata hubs such as IMDB [11] can be used to extend our approach in the movies domain as well.

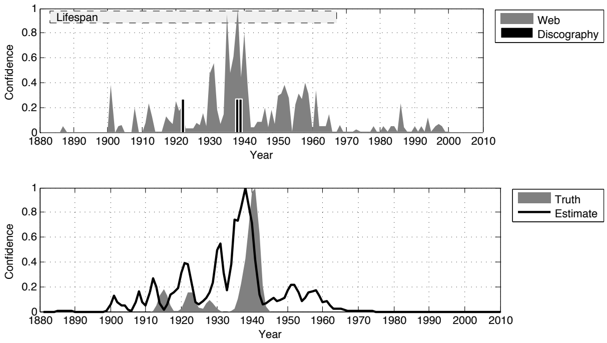

Figure 5. (Top) the input profiles, (bottom) the ground truth against the estimated artist productivity profile (APP) for August de Laat.

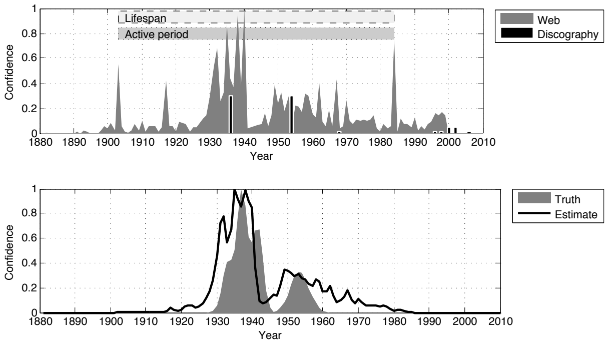

Figure 6. (Top) the input profiles, (bottom) the ground truth against the estimated APP for Bob Scholte.

8. ACKNOWLEDGMENTS

This research is supported by the NWO CATCH project COGITCH (640.005.004), and the FES project COMMIT/. The authors would like to thank Frans Wiering for his useful comments.

9. REFERENCES

[1] A. Kozbelt: "Performance time productivity and versatility estimates for 102 classical composers," *Psychology of Music*, Vol. 37, No. 1, pp. 25-46, 2009.

[2] A. Wilder: *American Popular Song: The Great Innovators, 1900 - 1950*, Oxford University Press, New York, 1990.

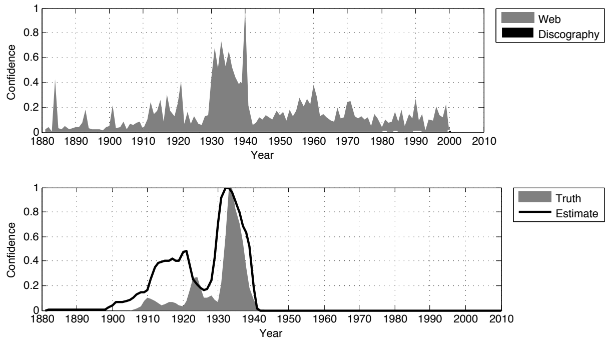

Figure 7. (Top) the input profiles, (bottom) the ground truth against the estimated APP for Louis Davids.

[3] B. Whitman and S. Lawrence: "Inferring descriptions and similarity for music from community metadata," *ICMC Conference Proceedings*, pp. 591-598, 2002.

[4] D. Bogdanov and P. Herrera: "Taking advantage of editorial metadata to recommend music," *CMMR Conference Proceedings*, pp. 618-632, 2012.

[5] D. K. Simonton: "Creative productivity: a predictive and explanatory model of career landmarks and trajectories," *Psycological Review*, Vol. 104, No. 1, pp. 66-89, 1997.

[6] F. Pachet, A. La Burthe, J. Aucouturier and A. Beurive: "Editorial metadata in electronic music distribution systems: between universalism and isolationism," *Journal of New Music Research*, Vol. 34, No. 2, pp. 173-184, 2005.

[7] J. Haitsma and T. Kalker: "A highly robust audio fingeprinting system," *ISMIR Conference Proceedings*, pp. 107-115, 2002.

[8] M. Schedl, C. Schiketanz and K. Seyerlehner: "Country of origin determination via web mining techiniques," *AdMIRe Proceedings*, pp. 1451-1456, 2010.

[9] M. Schedl and D. Hauger: "Mining microblogs to infer music artist similarity and cultural listening patterns," *WWW Conference Proceedings*, pp. 877-886, 2012.

[10] P. Knees, E. Pampalk and G. Widmer: "Artist classification with web-based data," *ISMIR Conference Proceedings*, pp. 517-524, 2004.

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

[12] T. Teitelbaum, P. Balenzuela, P. Cano, and J. M. Buldu: "Community structures and role detection in music networks," *Chaos journal*, Vol. 18, No. 4, pp. 043105, 2005.

[13] T. Bertin-Mahieux, D.P.W. Ellis, B. Whitman and P. Lamere: "The million song dataset," *ISMIR Conference Proceedings*, pp. 591-596, 2011.

[11] www.imdb.com