COUNTRY OF ORIGIN DETERMINATION VIA WEB MINING TECHNIQUES

Markus Schedl, Cornelia Schiketanz, Klaus Seyerlehner

Department of Computational Perception
Johannes Kepler University
Linz, Austria

markus.schedl@jku.at, music@jku.at, klaus.seyerlehner@jku.at

ABSTRACT

The origin of a music artist or a band is an important kind of musical meta-data as it usually influences his/her/its music. In this paper, we propose three approaches to automatically determine the country of origin of a person or institution, which we apply to music artists and bands. The first approach investigates estimates of *page counts* returned for specific queries to Web search engines. The second approach uses *term weighting functions* for country-specific terms that occur on the top-ranked Web pages of an artist. The third approach applies to Web pages *text distance measures* between country-specific terms and key terms related to the concept or origin. We further present a thorough evaluation of the approaches taking into consideration different refinements. We show that we are able to outperform the first, nevertheless recent, approach to determine the origin of a music artist.

Keywords— Web Mining, Country of Origin Detection, Term Weighting, Music Information Retrieval, Evaluation

1. INTRODUCTION

In times of steadily growing sizes of digital music collections – both commercial and private – the availability of various kinds of meta-data, such as line-up of a band, record release year, images of album covers, or artist biographies, are crucial to the music distributor in order to obtain a decisive advantage over its competitors. Likewise, for the private music aficionado, such meta-data is valuable as it enables, for example, browsing and filtering according to meta-data properties, semi-automated playlist generation, or clustering of music pieces with respect to meta-data attributes.

Some kinds of meta-data are frequently offered by record companies, music information systems such as *last.fm*[1] or *allmusic.com*[2], or enterprises specialized on the maintainance of meta-data catalogues such as *Gracenote*[3]. Since manually collecting meta-data is a very laborious and time-consuming task, and meta-data catalogues are usually not freely available, methods to automatically mine meta-data from the Web are important in the research area of *music information research*, especially in music information extraction (IE) and information retrieval (IR).

In the following, we use the term "artist" to refer to both individual musicians and music bands. The origin of an artist is obviously an interesting aspect of his or her life since it plays an important role as a semantic component in the understanding of the artist's context. For example, an artist's geographic and cultural context, political background, or song lyrics are likely strongly related to his or her origin.

2. RELATED WORK

To the best of our knowledge the only scientific work that also aims at automatically determining the origin of an artist is [1]. Govaerts and Duval's approach differ considerably from ours in that they rely on selected Web sites and services, such as *last.fm*, *Wikipedia*[4], and *Freebase*[5], instead of using potentially the whole Web for information extraction. They extract artist biographies from *Wikipedia* and propose three heuristics to determine the artist's country of origin using the occurrences of country names in these biographies. For evaluation they use a set of more than 11 000 artists from *Aristo Music*[6], which has been manually annotated by music experts. However, this set is very unevenly distributed with respect to continents since more than 95% of the artists originate from Europe or North America. The most likely reason for this is the commercial orientation of *Aristo Music*. We would also have liked to compare our approaches to that of Govaerts and Duval. Unfortunately, the *Aristo Music* data set used in [1] is not publicly available. As for Govaerts and Duval's results, they report that they were able to determine the origin for 59% of the test set, by at least one of the analyzed methods. A comparison among the three data sources showed that *Wikipedia* performed best with 56% coverage[7], *Freebase* performed second best with 26% coverage, followed by *last.fm* with only 7% coverage. Accuracy values varied between 70% (*Wikipedia*) and 90% (*last.fm* and *Freebase*).

3. DETERMINING AN ARTIST'S ORIGIN

The concept of "origin" is not unambiguously defined in literature, it rather depends on the context of its usage. In this paper, we define the "country of origin" of an artist as the country in

This research is supported by the *Austrian Fonds zur Förderung der Wissenschaftlichen Forschung* (FWF) in projects L511-N15 and Z159.

[1] http://last.fm (January 2010)

[2] http://www.allmusic.com (January 2010)

[3] http://www.gracenote.com (February 2008)

[4] http://www.wikipedia.org (December 2007)

[5] http://www.freebase.com (January 2010)

[6] http://www.aristomusic.com (January 2010)

[7] The coverage determines the percentage of artists for which a related Web page was found in the respective data source. It equals the concept of "recall" commonly used in the IR literature.

978-1-4244-7493-6/10/$26.00 ©2010 IEEE

which either the performer or musician was born or the band was founded. This definition sometimes led to interesting insights. For example, *Farrokh Bulsara*, also known as *Freddie Mercury*, was born in Zanzibar, United Republic of Tanzania. However, he relocated to the United Kingdom at the age of 17, where he later became world famous as co-founder of the band *Queen*. Mercury's country of origin is nevertheless Tanzania, whereas Queen's is the UK, where the band was founded by Mercury, Brian May, and Roger Taylor in April 1970.[8] This illustrative example is intended to highlight the problem of determining the country of origin in cases where the main country of musical activity differs from the place of birth.

We propose three different IE approaches to determine the origin of an artist: The first relies solely on the estimate of an artist's number of Web pages that contain the country term. We henceforth call this approach *Page Counts Approach*. The second approach applies term weighting measures commonly used in text-based information extraction and retrieval research, for example, [2, 3], to the retrieved Web pages. We will denote this strategy *Term Weighting Approach* in the following. The third approach uses heuristics based on the text distance between country names and key terms in the retrieved Web pages. We will refer to this method as *Text Distance Approach* in the following.

3.1. Web Page Retrieval

Regardless of the employed IE technique, the first step in our country-of-origin-prediction approach is to identify Web pages related to the artist under consideration, for example, fan pages, biographies, album reviews, track lists, or sale offers for albums or songs. This Web page selection can be carried out either by using a focused crawler or by relying on Web search engines. We follow the second approach here. Automatically querying a Web search engine to determine pages related to a specific topic is a common and intuitive task, which is nevertheless frequently performed in IE research. Examples in the music domain can be found in [4, 5], whereas [6, 7, 8] apply this technique in a more general context. Although this approach seems to be straightforward, it is prone to a major category of error: When searching for artist names that equal common speech words, usually a lot of irrelevant pages are returned.[9] Hence, the main challenge when using queries to a search engine for Web page selection is to restrict the search results to pages related to the desired artist. This problem is commonly addressed by enhancing the search query for the artist name with additional keywords. In the context of music information research, Whitman and Lawrence [4] proposed to confine the search by the keywords "music" and "review" in order to direct it towards album reviews. The resulting query scheme was successfully applied for genre classification tasks, e.g., [9]. To gather general, music-related Web pages, the scheme `"`*`artist`*`" music` usually represents a good trade-off between coverage and false positives. Hence, we used it for the paper at hand.

We first query *Google*'s search engine to retrieve up to the top 100 URLs for the artist for which the origin is to be determined. We then fetch the Web content available at these URLs. Subsequently, we create a *full inverted index*, also known as *world-level index*, [10] using a modified version of the open source indexer *Lucene Java*[10]. The resulting index is then taken as input to the IE approaches described in the following.

3.2. Page Counts Approach

This simple approach to determine an artist's origin extends the work presented in [5, 11]. The basic idea is to use a search engine's number of indexed Web pages for a given query, a count usually referred to as *page count*. Since these page counts are, however, only rough estimates of the real number of crawled Web pages related to the query, the results tend to be not very accurate. Nevertheless, for the purpose of classifying artists into genres [5, 11] and for classifying instances according to a given ontology as well as for learning sub- and superconcept relations [6, 7], this method yielded respectable results.

Using the search engine's API or issuing HTTP requests to the search engine and subsequently retrieving the resulting page count values for all ⟨artist, country⟩ tuples is the core component of this approach. To avoid excessive bandwidth consumption, however, we restrict the number of search results to be transmitted to the smallest value (this is usually one result). Since we are only interested in the page count estimates, this restriction effectively reduces network traffic without effecting the results. In our experiments we use *Google*'s search engine as it proved to outperform *Yahoo!*[11] and *MSN Search*[12], cf. [12, 9].[13] To alleviate the problem with artist names that equal common speech words, we apply the query scheme `"`*`artist`*`" "`*`country`*`" music`. From the resulting page count estimates of all ⟨artist, country⟩ tuples, we create an artist-country-matrix and eventually predict for each artist the country with the highest score.

3.3. Term Weighing Approaches

The standard procedure in text-based IR is to apply the *bag-of-words* model [13] to the documents under consideration. This model describes each document of a corpus by the contained words, irrespective of their ordering within the document, thus ignoring any structure or grammar rules. Words may also be generalized to *terms* by considering sequences of n consecutive words, so-called *n-grams*.

Using this bag-of-words representation of a document d, each term t is usually assigned a weight $w_{t,d}$ that reflects t's importance for document d. Integrating the weights of all terms for a specific document d yields a *feature vector* that describes d. Applying this procedure to the whole corpus of documents, each document can be described as a *term weight vector* and can be thought of as a representation of its weight vector in a vector space. The model underlying such a representation is often referred to as the *vector space model* and is fundamental in IR and IE research. It was originally described in [14].

If the aim is to describe music artists via content found on related Web pages, the complete set of pages retrieved for a particular artist a is often aggregated to a *virtual document* of a.

[8] `http://www.last.fm/music/Queen` (January 2010)

[9] In the music domain typical artists that cause such problems are *Bush*, *Prince*, *Kiss*, and *Porn*.

[10] `http://lucene.apache.org` (January 2008)

[11] `http://www.yahoo.com` (November 2008)

[12] `http://www.msn.com` (November 2008)

[13] In the meantime *Yahoo!* and *MSN Search* have agreed on merging their Web search knowledge and call the resulting search engine *bing* `http://www.bing.com` (January 2010).

Applying a term weighting measure $w_{t,a}$ to the virtual documents, each artist is described by a *term profile*, i.e., a vector space representation. Using such virtual documents seems reasonable since the subject of interest in Web-based music information retrieval is commonly the artist. Coping with small or empty pages is further facilitated if they are part of a larger virtual document.

For the work at hand, we used the following term weighting measures, since they are well founded in IR research, cf. [15, 16, 17].

Document frequency: $df_{t,a}$ is the total number of Web pages retrieved for artist a on which term t occurs at least once.[14]

Term frequency: $tf_{t,a}$ is the total number of occurrences of term t in the virtual document of a.

Term frequency·inverse document frequency: The basic idea of the $tf \cdot idf_{t,a}$ measure is to increase the weight of t if t occurs frequently in a virtual document of a, while at the same time decrease t's weight if t occurs in a large number of documents in the whole corpus, and is thus not very discriminative for a. We investigated two variants of the $tf \cdot idf$ measure since first experiments with a formulation that yielded good results for artist-to-genre classification [9] performed weakly for the task tackled in this paper. We refer to the formulation from [9] as $tf \cdot idf^1$:

$$tf \cdot idf^1_{t,a} = \begin{cases} (1 + \log_2 tf_{t,a}) \cdot \log_2 \frac{n}{df_t} & \text{if } tf_{t,a} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Experiments with slightly varying formulations eventually yielded variant $tf \cdot idf^2$:

$$tf \cdot idf^2_{t,a} = \ln\left(1 + tf_{t,a}\right) \cdot \ln\left(1 + \frac{n}{df_t}\right)$$

In both formulations n equals the total number of Web pages retrieved, and df_t is the total number of Web pages containing term t. Using the set of country names C as input, we calculate the weight for all terms $t \in C$ applying each term weighting function. Predicting the country for an artist is then simple performed by selecting the most important country term as determined by the term weighting measure.

3.4. Text Distance Approaches

As a third category of approaches, we tested several heuristics that make use of *text distances* between key terms K and country names C. The key terms comprise words like "born", "founded", "origin", and "country".

As text distance measure we use the *difference between the character offsets* of terms from C and K in a's Web pages. Using these differences, we build a model of a's most likely country of origin. The core part of this model integrates two different functions: first, a *distance measure* on the *document-level* (dlf) to determine the distances within a Web page of a; second, an *aggregation function* (af) to combine the document-level-distances for all pages retrieved for a. The choice of these two functions is vital to the quality of the prediction. For the evaluation experiments, we use the following scheme to describe a setting: $\{\mathit{key}_1, \cdots, \mathit{key}_n\}$, dlf, af. For example, in the setting {born, founded}, avg, min the list of key terms comprise the words "born" and "founded", the dlf distance measure is the arithmetic mean of the distances between country names and key terms, and the minimum is used as aggregation function.

3.5. Synonym Lists for Countries and Nationalities

Analyzing frequent errors, we gained the insight that certain countries often tend to be erroneously predicted. For example, the "United States" are frequently incorrectly predicted for two reasons: First, the term "United States" is not only used to denote the "United States of America", therefore resulting in unjustified higher term weights. Second, even if the meaning is correct, the occurrence of "United States" on a Web page may also refer to various relations other than to the country of origin. To improve accuracy by mitigating these two problems, we investigated the use of *synonyms for country names and related terms*, e.g., the respective nationalities.[15] To this end, we gathered synonyms for countries and nationalities from *Thesaurus.com*[16]. Each country is therefore described not only by its name, but also by related terms.

We aggregate the synonyms for a country by calculating the *arithmetic mean* of the respective scoring measure (term weight or text distance). We also tried using minimum, maximum, and median for aggregation. However, taking the minimum leads to an underestimate of the importance of some countries. For example, "Land of Opportunity" is considered a synonym for the USA, but is seldom used in Web pages to refer to this country. Using instead the maximum function for aggregation causes serious distortions for synonyms that equal common speech words. For example, "US" as a synonym for the United States obviously yields many erroneous predictions for the USA since it is a word too frequently used to refer to the pronoun "us". The median yielded similar, but slightly worse results than the arithmetic mean.

4. EVALUATION

4.1. Test Collection

Since there exists no standardized data set for this kind of task, we had to build one on our own. To this end, we manually gathered 578 artists and their country of origin from all over the world from *Wikipedia*, *last.fm*, and *allmusic.com*.[17] We included artists from 69 distinct countries. In total 50 967 Web pages were gathered applying the Web page retrieval procedure described in Section 3.

4.2. Coverage and Precision

We were foremost interested in the quantity of artists for which a country of origin can be determined and in the quality of the prediction. Hence we investigated *coverage* (or *recall*) and *precision* of the proposed approaches, coverage being defined as the percentage of artists for which a country of origin could be determined, precision being defined as the number of artists

[14] In this case, the single Web pages retrieved for a are considered, instead of a's aggregated virtual document.

[15] The list of synonyms is available at http://www.cp.jku.at/people/schedl/music/countries_syn.txt.

[16] http://thesaurus.reference.com (January 2010)

[17] The data set can be downloaded from http://www.cp.jku.at/people/schedl/music/C578a_country.txt.

Approach	**C (%)**	**P(%)**	**F**
Page counts			
Google	**100**	**23.18**	**37.64**
Term weighting (without synonyms)			
df	100	65.57	79.21
tf	**100**	**68.86**	**81.56**
$tf \cdot idf^1$	100	57.96	73.38
$tf \cdot idf^2$	100	63.49	77.67
Term weighting (with synonyms)			
df	100	66.09	79.58
tf	**100**	**70.76**	**82.88**
$tf \cdot idf^1$	100	54.50	70.55
$tf \cdot idf^2$	100	59.34	74.48
Text distance (without synonyms)			
{born}, min, min	100	34.08	50.84
{born, founded}, min, min	**100**	**37.20**	**54.22**
{born}, avg, min	100	14.19	24.85
{born, founded}, avg, min	100	14.19	24.85
Text distance (with synonyms)			
{born}, min, min	100	29.41	45.45
{born, founded}, min, min	**100**	**32.53**	**49.09**
{born}, avg, min	100	12.11	21.60
{born, founded}, avg, min	100	12.46	22.15

Table 1. Evaluation results.

Approach	**C (%)**	**P(%)**	**F**
$last.fm\_origin$	7.19	89.58	13.13
$freebase\_origin$	21.37	90.85	34.60
$freebase\_most\_freq$	26.20	**91.60**	40.75
$wikipedia\_most\_freq$	55.76	64.63	59.87
$combined\ method$	**59.12**	77.09	**66.92**

Table 2. Evaluation results from Govaerts and Duval [1].

Approach	**with**	**w/o**	**z**	**sgn**
df	66.09	65.57	-0.38	
tf	70.76	68.86	-1.39	
$tf \cdot idf^1$	54.50	57.96	-2.09	*
$tf \cdot idf^2$	59.34	63.49	-2.72	*
{born}, min, min	29.41	33.91	-2.64	*
{born, founded}, min, min	32.53	37.20	-2.64	*
{born}, avg, min	12.11	14.19	-2.06	
{born, founded}, avg, min	12.46	14.19	-1.89	*

Table 3. A statistical comparison of the approaches with and without the use of synonyms.

whose country was correctly predicted divided by the number of artists for which a prediction was made. As an aggregate measure of precision and recall, we further report the *F-measure* [18], which is the weighted harmonic mean of precision and recall.

4.2.1. *Analysis and Discussion*

Table 1 shows *coverage*, *precision*, and *F-measure* for each category of approaches and a selection of parameter settings within these categories. The best performing setup is highlighted within each category of methods.

Table 2 reproduces the results obtained by Govaerts and Duval in [1]. Unfortunately, the authors used a proprietary test collection that is not publicly available. Therefore, their results are only roughly comparable to ours. Further note that even though Govaerts and Duval used a 11 000-artist-collection, their best results (which are illustrated in Table 2) were achieved on a subset of 3 000 artists.

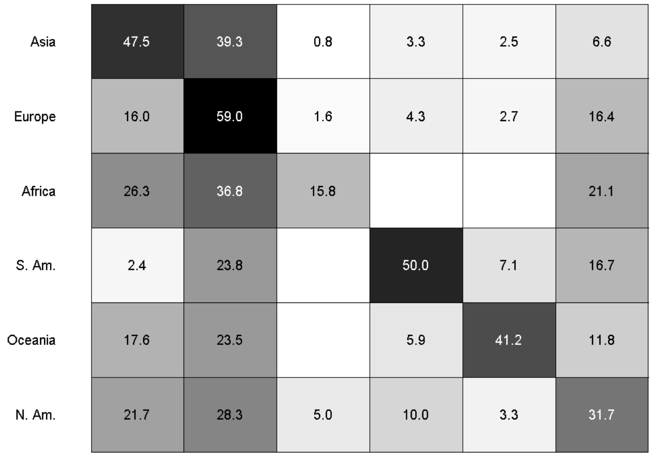
Compared to [1] (Table 2), our approaches (Table 1) reach a higher coverage (100%). This is no surprise as our approaches may incorporate, at least in theory, the whole Web. In contrast, the precision is usually smaller for our general Web-based approaches. This seems reasonable as Govaerts and Duval use very specific Web sites to extract information (*Wikipedia*, *Freebase*, and *last.fm*). Taking a look at the F-mesaure we see that our term weighting approaches outperform all methods proposed in [1]. Even our best performing text distance measures, which in general scored much worse than the term weighting approaches, perform similar to the best performing single measures (not the combined method) from [1]. As part of future work, we would like to combine the high precision of Govaert and Duval's methods with the high coverage of our approaches by integrating the different data sources.

Taking a closer look at the individual categories of approaches, large differences become apparent. The simple page counts approach seems to be too simple to capture the semantic category of country of origin. In contrast, applying term weighting functions to a selection of top-ranked, artist-related Web pages yields the best results. An interesting finding in this context is that the *tf* and *df* measures outperform the $tf \cdot idf$-based measures. $tf \cdot idf$ is the standard approach in text-based IR, but underperforms in this specific IE task. This is likely a result of $tf \cdot idf$'s penalization of terms that occur within a large number of documents. In standard IR the idf factor is used to demote words that do not bear much discriminative power as they appear in many documents. Suppressing such terms does make sense in most IR tasks, which aim at finding documents specific to a query. In our IE task, in contrast, general and popular terms should not be given less weight. This result is in line with the findings of [19].

Interestingly, the text distance approaches that tackle the problem in a very specific manner, and which we therefore expected to outperform the more general term weighting approaches, performed worse. A possible explanation is that we may have chosen the wrong terms for the set of key terms, in which case the country names simply do not occur too close to the chosen anchor terms. Also the use of synonyms did not improve results for the text distance approaches, in contrast to the term weighting approaches. An explanation is suggested in the following subsection.

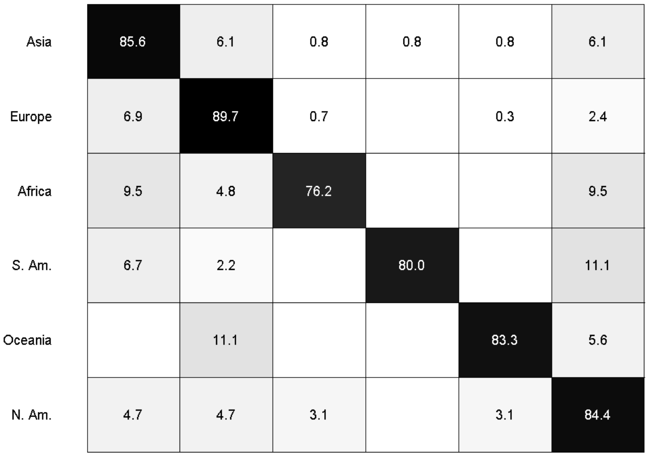
4.2.2. *Statistical Significance Tests*

We were first interested in whether using synonyms significantly impacts the obtained results of the various approaches. For each pair of approaches, with synonyms and without synonyms, we test the equality of the groups' medians using the *Wilcoxon signed-rank test* [20]. In Table 3 all significant differences are marked. There exists a significant difference for



Asia	47.5	39.3	0.8	3.3	2.5	6.6
Europe	16.0	59.0	1.6	4.3	2.7	16.4
Africa	26.3	36.8	15.8			21.1
S. Am.	2.4	23.8		50.0	7.1	16.7
Oceania	17.6	23.5		5.9	41.2	11.8
N. Am.	21.7	28.3	5.0	10.0	3.3	31.7

Fig. 1. Confusion matrix for the page counts approach ($Google$).



Asia	85.6	6.1	0.8	0.8	0.8	6.1
Europe	6.9	89.7	0.7		0.3	2.4
Africa	9.5	4.8	76.2			9.5
S. Am.	6.7	2.2		80.0		11.1
Oceania		11.1			83.3	5.6
N. Am.	4.7	4.7	3.1		3.1	84.4

Fig. 2. Confusion matrix for the best term weighting approach (*tf* with synonyms).

$tf \cdot idf$-based approaches. Furthermore, three of the approaches based on text distances perform significantly worse if synonyms are used. This seems reasonable since for this group of approaches, the use of ambiguous synonyms, such as “US”, “Johnny”, or “Sam”, causes a high number of incorrect predictions.

It is further interesting to investigate if there are significant differences between the approaches in each group of Table 1. To identify significant differences within the groups, *Friedman's two-way analysis of variance* [21] is used. The test revealed highly significant differences between all categories of approaches. As post-hoc test to analyze which settings significantly differ within their category, the *Wilcoxon signed-rank test* [20] is used again, and the significance level is adjusted for multiple comparisons. All approaches that significantly differ from the best performing approach in each group are marked in italics in Table 1. Except for the term weighting group without synonyms, where no significant difference between *df* and *tf* could be determined, the performance of the best approach is always significantly different from all others.

4.3. Confusions

We performed *confusion analysis* to investigate which countries are often incorrectly predicted. To this end, we aggregated the countries to continents, as a detailed country-wise analysis would have been beyond this paper's scope. Figures 1, 2, and 3 depict confusion matrices for the best performing setups within each category of approaches, respectively page counts, term weighting, text distance. Along the ordinate the correct continents are illustrated, whereas the predicted continents can be found along the abscissa. Considering Figure 1, for example, 47.5% of Asian-born or -founded artists are correclty classified as Asian by the page counts approach. However, 39.3% of the Asian artists are incorrectly classified as originating from Europe.

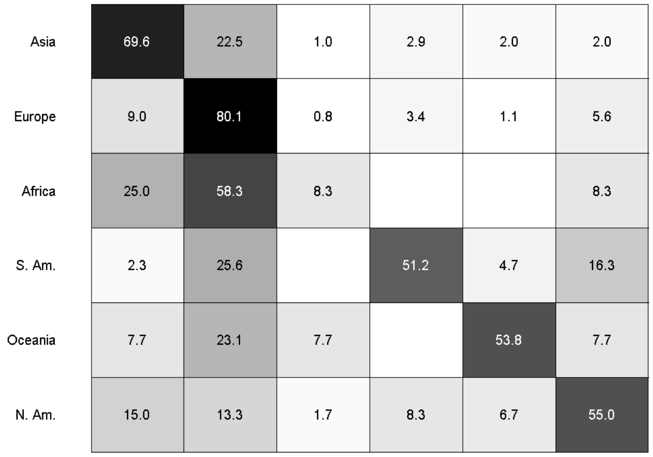
4.3.1. Analysis and Discussion

The most obvious finding from the confusion analysis is that artists from other parts of the world are often incorrectly classified as originating from Europe or North America. This becomes particularly evident in the case of the simple page counts approach (Figure 1). More than 20 percent of all artists from any continent other than Europe are misclassified as being European. Furthermore, except for Asians, more than 10 percent of all artists not from North America are misclassified as North Americans. Using the simple page counts approach seems to introduce a strong bias towards those continents where Web coverage is highest.

Analyzing which continents suffer the most from artists that originate from there, but are wrongly classified to originate from other continents, this is the case for Africa. Only 15.8%, 76.2%, and 8.3% of artists originating from Africa are correctly classified using the page counts approach, the term weighting approach, and the text distance approach, respectively. For the term weighting approach, this number is remarkably high, nevertheless the worst among all continents.

5. CONCLUSION AND FUTURE WORK

We presented three approaches to automatically determine the country of origin of a person or institution via Web mining techniques, and we applied these approaches to the problem of finding the origin of a music artist or band. The first approach investigates estimates of *page counts* returned for specific queries to Web search engines. It was shown that this approach is too simple to correctly predict the country of origin in most cases. The second approach applies *term weighting functions* for country-specific terms that occur on the top-ranked Web pages of an artist. Using the term frequency as weighting measure, we achieved the best results, both in terms of coverage and precision. The third approach makes use of *text distance measures* between country-specific terms and origin-related key terms on the retrieved Web pages. It yielded worse results than the term weighting approach. Even though the proposed approaches are



Asia	69.6	22.5	1.0	2.9	2.0	2.0
Europe	9.0	80.1	0.8	3.4	1.1	5.6
Africa	25.0	58.3	8.3			8.3
S. Am.	2.3	25.6		51.2	4.7	16.3
Oceania	7.7	23.1	7.7		53.8	7.7
N. Am.	15.0	13.3	1.7	8.3	6.7	55.0

Fig. 3. Confusion matrix for the best text distance approach ($\{$*born, founded*$\}$*, min, min* without synonyms).

of a rather simple nature, we were able to outperform earlier work [1].

We further showed that using a list of synonyms for country-specific terms can improve precision and F-measure for the term weighting approaches, but worsens the results for the text distance approaches.

A confusion analysis on the level of continents revealed that artists from all over the world are often incorrectly classified as originating from Europe or North America.

In the future, we will further investigate the use of *natural language processing* (NLP) techniques to improve precision. Moreover, we would like to generalize our approach to a *broader spectrum of application domains*. The methods proposed here can easily be applied to determine the country of origin of any person (or institution) that is popular enough to be mentioned on a considerable number of Web pages. However, not all Web pages contain reliable, high-quality information. Another part of future work will therefore consist of determining the pages one can rely on, using approaches to estimate the reputation of a Web page, such as [22].

6. REFERENCES

[1] Sten Govaerts and Erik Duval, "A Web-based Approach to Determine the Origin of an Artist," in *Proc. of ISMIR*, Kobe, Japan, 2009.

[2] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley, 1999.

[3] *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.

[4] Brian Whitman and Steve Lawrence, "Inferring Descriptions and Similarity for Music from Community Metadata," in *Proc. of ICMC*, Göteborg, Sweden, 2002.

[5] Gijs Geleijnse and Jan Korst, "Web-based Artist Categorization," in *Proc. of ISMIR*, Victoria, Canada, 2006.

[6] Philipp Cimiano, Siegfried Handschuh, and Steffen Staab, "Towards the Self-Annotating Web," in *Proc. of ACM WWW*, New York, NY, USA, 2004.

[7] Philipp Cimiano and Steffen Staab, "Learning by Googling," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 2, 2004.

[8] Peter Knees, Tim Pohle, Markus Schedl, and Gerhard Widmer, "A Music Search Engine Built upon Audio-based and Web-based Similarity Measures," in *Proc. of ACM SIGIR*, Amsterdam, the Netherlands, 2007.

[9] Peter Knees, Elias Pampalk, and Gerhard Widmer, "Artist Classification with Web-based Data," in *Proc. of ISMIR*, Barcelona, Spain, 2004.

[10] Justin Zobel and Alistair Moffat, "Inverted Files for Text Search Engines," *ACM Computing Surveys*.

[11] Markus Schedl, Tim Pohle, Peter Knees, and Gerhard Widmer, "Assigning and Visualizing Music Genres by Web-based Co-Occurrence Analysis," in *Proc. of ISMIR*, Victoria, Canada, 2006.

[12] Sten Govaerts, Nik Corthaut, and Erik Duval, "Using Search Engine for Classification: Does It Still Work?," in *Proc. of AdMIRe*, San Diego, CA, USA, 2009.

[13] Hans Peter Luhn, "A Statistical Approach to Mechanized Encoding and Searching of Literary Information," *IBM Journal*, 1957.

[14] Gerard Salton, A. Wong, and C. S. Yang, "A Vector Space Model for Automatic Indexing," *Communications of the ACM*, vol. 18, no. 11, 1975.

[15] Justin Zobel and Alistair Moffat, "Exploring the Similarity Space," *ACM SIGIR Forum*, vol. 32, 1998.

[16] Gerard Salton and Christopher Buckley, "Term-weighting Approaches in Automatic Text Retrieval," *Information Processing and Management*, vol. 24, 1988.

[17] Franca Debole and Fabrizio Sebastiani, "Supervised Term Weighting for Automated Text Categorization," in *Proc. of ACM SAC*, Melbourne, FL, USA, 2003.

[18] C. J. van Rijsbergen, *Information Retrieval*, Butterworths, 2nd edition.

[19] Markus Schedl and Tim Pohle, "Enlightening the Sun: A User Interface to Explore Music Artists via Multimedia Content," *Multimedia Tools and Applications: Special Issue on Semantic and Digital Media Technologies*, October 2009.

[20] Frank Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, vol. 1, 1945.

[21] Milton Friedman, "A Comparison of Alternative Tests of Significance for the Problem of m Rankings," *The Annals of Mathematical Statistics*, vol. 11, 1940.

[22] Davood Rafiei and Alberto O. Mendelzon, "What is this Page Known for? Computing Web Page Reputations," *Computer Networks*, vol. 33, no. 1-6, 2000.