Three Web-based Heuristics to Determine a Person's or Institution's Country of Origin

Markus Schedl [1]
markus.schedl@jku.at

Klaus Seyerlehner [1]
klaus.seyerlehner@jku.at

Dominik Schnitzer [1,2]
dominik.schnitzer@ofai.at

Gerhard Widmer [1,2]
gerhard.widmer@jku.at

Cornelia Schiketanz [1]
music@jku.at

[1] Dept. of Computational Perception
Johannes Kepler University (JKU)
Linz, Austria

[2] Austrian Research Institute for
Artificial Intelligence (OFAI)
Vienna, Austria

ABSTRACT

We propose three heuristics to determine the *country of origin* of a person or institution via *text-based IE from the Web*. We evaluate all methods on a collection of *music artists and bands*, and show that some heuristics outperform earlier work on the topic by terms of coverage, while retaining similar precision levels. We further investigate an extension using country-specific *synonym lists*.

Categories and Subject Descriptors: H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing I.7.m [Document and Text Processing]: Web Mining

General Terms: Algorithms, Measurement

Keywords: information extraction, country of origin detection, term weighting, music information research, evaluation

1. INTRODUCTION AND CONTEXT

The country of origin of a person or institution represents an interesting aspect of his/her/its life/existence. It plays a vital role in understanding a person's background and context. As a *semantic descriptor*, the country of origin of a music artist or band, both of which we will refer to as "artist" in the following, can be used to *query music collections* based on learned associations between acoustic features and textual features – cf. [2, 6]

The "country of origin" of an artist is defined as the country where he or she was born, or the band was founded. What makes this task a challenge is foremost that the origin is neither always unambiguous, nor well-known. Consider, for example, *Farrokh Bulsara*, later known as *Freddie Mercury*. He was born in Zanzibar, Tanzania. However, he relocated to the UK 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. This example highlights the problem of determining the origin in cases where the main country of musical activity differs from the place of birth.

Earlier work on predicting the origin of a music artist mainly consists of [3, 4]. In contrast to our approaches that may – at least in theory – use the whole Web as data source, Govaerts and Duval focus on specific Web sites, such as *last.fm*, *Wikipedia*, and *Freebase*, and apply heuristic functions on their textual content. For evaluation they use a commercial, non publicly available set of artists, which has been manually annotated by music experts.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

2. HEURISTICS

The first and simplest heuristic investigates estimates of search engine's *page counts* for queries containing the artist to be classified and the country name. We use *Google*'s index since this engine has already proven to yield respectable results for artist-to-genre classification tasks based on weighted term features – cf. [5]. To mitigate the distortions arising from artist names that equal common speech words[1], we employ the query scheme `"artist" "country" music`. We finally predict the country whose page count is highest for a given artist, formally $max_c\ pc(artist,\ c)\ \forall c \in C$, where C is the set of country names.

The second approach applies *term weighting functions* [1] to the textual content of the 100 top-ranked Web pages retrieved from Google's index for the person under consideration. We use the following term weighting measures since they are well founded in IR research: *document frequency (df)*, *term frequency (tf)*, and $tf \cdot idf$ of the country terms in the set of artist-related Web pages. We conducted experiments with various $tf \cdot idf$ variants and found that the following seems to be suited best for this particular task:

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

In this formulation $tf_{t,a}$ denotes the number of occurrences of term t in the 100-page-set retrieved for artist a, df_t represents the number of pages where t occurs among the complete set of all Web pages retrieved, and n is the total number of pages retrieved. The origin of an artist a is then determined by predicting the country whose name ranks highest with respect to the employed term weighting function.

The third approach uses *text distance measures* between country names and *origin-related key terms*, such as "born", "founded", "origin", or "country", on the set of top-ranked Web pages. Based on the *offset at character-level* between the country terms and the origin-related key terms in a's pages, we build a model of a's most likely country of origin. The core part of this model integrates two different functions: a *distance measure* on the *intra-page-level* (ipl) to determine the distances within a Web page of a, and an *aggregation function* (af) to combine the ipl-distances for all pages retrieved for a. The choice of these two functions is

[1]Examples of such artists are "Bush", "Prince", and "Kiss".

vital to the quality of the prediction. For the evaluation experiments described next, we use the following scheme to describe a setting: $\{key_1, \cdots, key_n\}$, *ipl*, *af*.

Using Country Synonyms.

We further looked into using synonyms for countries and nationalities, extracted from *Thesaurus.com*. A complete list of the used mappings $country \mapsto \langle syn_1, \cdots, syn_n \rangle$ is available.[2] This step is motivated by the fact that certain countries, such as the "United States" (of America), are often wrongly predicted due to their *ambiguity*, and *abundant presence* on the Web.

3. EVALUATION AND DISCUSSION

Since there exists no standardized data set for this kind of task and we did not have access to the one used in [3], we manually extracted 578 artists and their country of origin from sources such as *allmusic.com*, *last.fm*, and *Wikipedia*.[3] We included artists from 69 distinct countries of the world.

Table 1 shows the evaluation results in terms of *coverage* (or *recall*), *precision*, and *F-measure* [7]. Coverage denotes the share of artists for which a country could be determined, precision is the share of artists whose origin is correctly predicted among the number of artists for which a prediction was made, and the F-measure aggregates precision and recall via the weighted harmonic mean. The best-performing approaches within each category are printed in bold.

The *page counts* approach seems to be too simple to capture the country of origin correctly. The *term weighting* approaches yield overall the best results. Interestingly, *tf* and *df* measures outperform $tf \cdot idf$. Even though $tf \cdot idf$ is the standard approach in text-based IR, it underperforms in this specific IE task. This is likely a result of $tf \cdot idf$'s penalization of terms that occur in a large number of documents. Suppressing such terms does make sense in most IR tasks. In our IE task, however, general and popular terms should not be given less weight. The *text distance* approach performs worse than expected. The reason for this bad performance may be an unfavorable set of key terms. We will investigate this as part of future work.

Table 2 shows the best evaluation results from Govaerts and Duval [3]. Comparing Tables 1 and 2, our approaches perform, in general, better with respect to coverage and F-measure. In terms of precision, the picture is more diversified. While Govaerts and Duval's combined method reaches about 77% precision (at a 59%-recall-level), our best method in terms of precision achieves about 71% (but at a 100%-recall-level).

Synonyms significantly impact the obtained results.

Employing the *Wilcoxon signed-rank test* on each pair of approaches (with and without synonyms) revealed significant difference for $tf \cdot idf$-based approaches. Furthermore, three of the approaches based on text distances perform significantly worse if synonyms are used. This may be explained by ambiguous synonyms, such as "US" or "Sam".

Significant differences between the three heuristics.

Friedman's two-way analysis of variance revealed highly significant differences between all categories of approaches. We further employed the post-hoc *Wilcoxon signed-rank test* to analyze which settings differ within their category. In Table 1 the settings that significantly differ from the best performing setting in each group are marked in italics. Except for the term weighting group without synonyms, where no significant difference between *df* and *tf* could be determined, the performance of the best setting is always significantly different from all others.

[2] http://www.cp.jku.at/people/schedl/music/countries_syn.txt

[3] http://www.cp.jku.at/people/schedl/music/C578a_country.txt

Approach	C (%)	P(%)	F
Page counts			
Google	**100.00**	**23.18**	**37.64**
Term weighting (without synonyms)			
df	100.00	65.57	79.21
tf	**100.00**	**68.86**	**81.56**
$tf \cdot idf$	100.00	63.49	77.67
Term weighting (with synonyms)			
df	100.00	66.09	79.58
tf	**100.00**	**70.76**	**82.88**
$tf \cdot idf$	100.00	59.34	74.48
Text distance (without synonyms)			
{*born*}, *min*, *min*	100.00	34.08	50.84
{born, founded}, min, min	**100.00**	**37.20**	**54.22**
{*born*}, *avg*, *min*	100.00	14.19	24.85
{*born, founded*}, *avg*, *min*	100.00	14.19	24.85
Text distance (with synonyms)			
{*born*}, *min*, *min*	100.00	29.41	45.45
{born, founded}, min, min	**100.00**	**32.53**	**49.09**
{*born*}, *avg*, *min*	100.00	12.11	21.60
{*born, founded*}, *avg*, *min*	100.00	12.46	22.15

Table 1: Evaluation results of our approaches.

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: Best evaluation results from [3].

4. CONCLUSIONS

We presented three parameterizable heuristics to determine the origin of a person or institution and applied these heuristics with different settings to a set of music artists and bands. We were able to outperform earlier work in terms of coverage and F-measure, while retaining precision levels. Future work will include refining our methods by combining them with *NLP* techniques or estimates of *Web page reputation*.

Acknowledgments

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

5. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[2] D. Turnbull et al. Towards Musical Query-by-Semantic Description using the CAL500 Data Set. In *Proc of 30th ACM SIGIR*, Jul 2007.

[3] S. Govaerts and E. Duval. A Web-based Approach to Determine the Origin of an Artist. In *Proc of 10th ISMIR*, Oct 2009.

[4] M. Schedl et al. Country of Origin Determination via Web Mining Technique. In *Proc of 2nd AdMIRe*, Jul 2010.

[5] P. Knees et al. Artist Classification with Web-based Data. In *Proc of 5th ISMIR*, Oct 2004.

[6] P. Knees et al. A Music Search Engine Built upon Audio-based and Web-based Similarity Measures. In *Proc of 30th ACM SIGIR*, Jul 2007.

[7] C. J. van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, London, UK, 2nd ed., 1979.