TOOL PLAY LIVE: DEALING WITH AMBIGUITY IN ARTIST SIMILARITY MINING FROM THE WEB

Gijs Geleijnse **Jan Korst**
Philips Research
Eindhoven (the Netherlands)
{gijs.geleijnse,jan.korst}@philips.com

ABSTRACT

As methods in artist similarity identification using Web Music Information Retrieval perform well on known evaluation sets, we investigate the application of such a method to a more realistic data set. We notice that ambiguous artist names lead to unsatisfying results. We present a simple, efficient and unsupervised method to deal with ambiguous artist names.

1 INTRODUCTION

Web Music Information Retrieval (Web-MIR) is a novel and promising field of research. Recently, excellent performances (with precision rates around 90%) are reported on artist genre classification and artist similarity identification using dedicated test sets [2, 1, 3].

In this work, we focus on the task of identifying and scoring artist similarities using web data. Although our method performs well on two commonly used test sets, the results on a more realistic collection of artists are less encouraging. Contrary to the two common benchmark sets for Web-MIR [2, 3], this set of artists contains ambiguous names of less famous artists. When querying for artist names, within the search results the ambiguous name *Nirvana* can be expected to refer to the band. However, the meaning of *Play* – also a band in our collection – is less obvious. We observe that such ambiguous artist names tend to be found similar to a large number of artists. Especially for lesser known artists, this leads to unsatisfying results. We therefore present an unsupervised method to deal with the phenomenon of ambiguous artist names, making the method more robust and suited for realistic tasks.

2 FINDING ARTIST SIMILARITIES

Using two efficient methods introduced in earlier work [1], we compute the measure of similarity $T(a, b)$ of artist a to another artist b using co-occurrences on the web. With PM, we extract artist names from search engine snippets after querying with relation patterns, where with DM we scan full documents for artist names. We use a set of 224 artists, equally divided over 14 genres [2] and a large set of 1995 artists divided over 9 genres [3]. We consider two artists to be similar, if they share a genre in the test set. Figure 1 shows the average precision of the similarity of the artists and their k-NN. We can conclude that the pattern based method gives good results and outperforms DM in both sets.

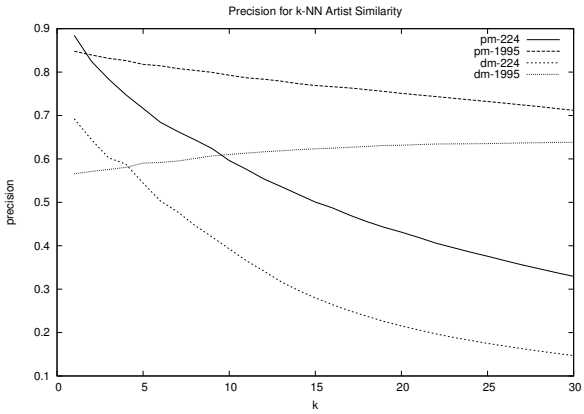

Figure 1. Precision for the sets of 224 and 1995 artists.

The experiments show that PM both outperforms DM and is less time consuming. Hence, we applied PM to a collection of 1732 artists, used in music recommender experiments. Since no ground truth is available for this collection of artist, we cannot evaluate the precision. We observe that the names that often not refer to the intended artist frequently occur amidst the most related artists. Although the queried expressions contain an artist name by construction, the retrieved snippets do not always handle musical artists and their relations.

3 DEALING WITH AMBIGUITY

Ideally, for each occurrence of an artist name in a text we want to observe whether the occurrence indeed reflects the intended artist. However, the automatic parsing of sentences and term recognition is troublesome as the snippets contain broken sentences and may be multilingual. Moreover if an artist name is identified as a subject or object within a sentence, then we still do not know whether the term indeed reflects the artist.

We therefore aim for an unsupervised method where we estimate the probability that a term a indeed reflects

the intended artist named a. If we know the probability $p(a)$ that a reflects the artist, we can redefine the artist similarity function T as follows.

$$T'(a,b) = \frac{\text{co}'(a,b)}{\sum_{c,c\neq b} \text{co}'(c,b)} \quad (1)$$

with

$$\text{co}'(a,b) = \text{co}(a,b) \cdot p(a) \cdot p(b) \quad (2)$$

Note that for $p(a) = p(b) = 1$, we have the baseline function [1] as applied in Section 2.

We use the *define* functionality in Google to obtain the number of senses of a term. For example, by querying `define: Tool`, we obtain a list of 31 definitions for the term *Tool*, collected from various online dictionaries and encyclopedias. This indicates that Tool is an ambiguous term. On the contrary, terms such as *Daft Punk*, *Fatboy Slim* and *Johannes Brahms* lead to precisely one definition. We define $n(a)$ to be the number of definitions for a to be returned by Google. If no definitions are returned, we consider $n(a)$ to be 1.

Based on the number of definitions $n(a)$ returned by Google, we investigate the following two alternatives to estimate $p(a)$.

linear. As we do not know anything about the distributions of the use of the definitions for term a, we estimate that each definition has a equal probability to be used and that only one of the definitions reflects the artist.

$$p_{\text{lin}}(a) = \frac{1}{n(a)} \quad (3)$$

sqrt. Especially for terms with many definitions, we observe some overlap between the definitions. Moreover, two distinct definitions can be close related. For example, *Red Hot Chili Peppers* is the name of a band and the name of their self-titled debut album. We therefore investigate a second method to estimate $p(a)$ by using the square root of the number of definitions found.

$$p_{\text{sqrt}}(a) = \frac{1}{\sqrt{n(a)}} \quad (4)$$

4 EXPERIMENTAL RESULTS

For both the sets of 224 and 1995 artists, we collected the numbers of definition for all the artist names. We recomputed the artist similarities using the linear approach (3) and the sqrt approach (4) and compared the two with the baseline as applied in Section 2. We present the results for the 1995 artists in Figure 2. For the set of 224 the performance of the methods using disambiguation is slightly less than that of the baseline approach. For the set of 1995 artists however, the results improve using either the *linear* or the *sqrt* approach. We note that contrary to the set of 224 artists, the 1995 set does contain 'difficult' ambiguous names such *Autograph*, *Gamma Ray* and *Hypocrisy*.

For the set of 1732 artists in our own collection, we compare the number of times that ambiguous artist names occur among the 5 nearest neighbors for the other artists

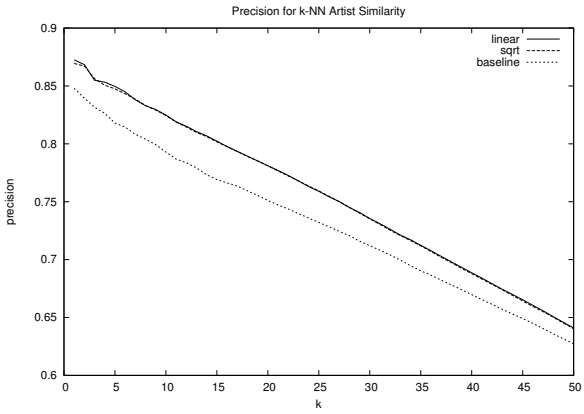

Figure 2. Precision for the sets of 1995 artists using the three ambiguity estimators.

(Table 1). We note that for the term *Juli* only one definition is found. Although the distribution of ambiguous names is quite different for p_{lin} and p_{sqrt}, we cannot draw conclusions which approach is better suited as currently no ground truth for artist similarity ranking is available.

Artist	baseline	p_{lin}	p_{sqrt}
Live	1227	1	54
Tool	1334	0	642
Fish	724	0	7
Juli	691	1251	1207

Table 1. Number of times an ambiguous artist name occurs among the top 5 nearest neighbors of the 1731 other artists.

5 CONCLUSIONS AND FUTURE WORK

We have shown that a method that performs well on a 'clean' set with few ambiguous names leads to unsatisfying results on a more realistic data set. Terms that do not have the intended artist as dominant meaning (e.g. *Boston*, *Play* and *Live*) are likely to be found similar to other artists. We observe this problem especially for lesser known artists, where the data collected is more sparse. We have shown that a simple, efficient and unsupervised method using the number of definitions of a term can compensate for this phenomenon.

6 REFERENCES

[1] G. Geleijnse and J. Korst. Web-based artist categorization. In *Proceedings of ISMIR'06*, pages 266 – 271, 2006.

[2] P. Knees, E. Pampalk, and G. Widmer. Artist classification with web-based data. In *Proceedings of ISMIR'04*, pages 517 – 524, 2004.

[3] M. Schedl, T. Pohle, P. Knees, and G. Widmer. Assigning and visualizing music genres by web-based co-occurrence analysis. In *Proceedings of ISMIR'06*, pages 260 – 265, 2006.