WEB-BASED DETECTION OF MUSIC BAND MEMBERS AND LINE-UP

Markus Schedl[1] **Gerhard Widmer**[1,2] **Tim Pohle**[1] **Klaus Seyerlehner**[1]
[1]Department of Computational Perception, Johannes Kepler University, Linz, Austria
[2]Austrian Research Institute for Artificial Intelligence, Vienna, Austria

ABSTRACT

We present first steps towards the automatic detection of music band members and instrumentation using web content mining techniques. To this end, we combine a named entity detection method with rule-based linguistic text analysis. We report on preliminary evaluation results and discuss limitations of the current method.

1 INTRODUCTION AND CONTEXT

Automatic extraction of textual information about music artists can be used, for example, to enrich music information systems, for automatic biography generation, to build relationship networks, or to define similarity measures between artists, a key concept in music information retrieval. Here, we present an approach to finding the members of a given music band and the respective instruments they play. In this preliminary work, we restrict instrument detection to the standard line-up of most Rock bands, i.e. we only check for singer(s), guitarist(s), bassist(s), drummer(s), and keyboardist(s).

2 METHODS

Basically, our approach comprises four steps: web retrieval, named entity detection, rule-based linguistic analysis, and rule selection.

Web Retrieval

Given a band name B, we use *Google* to retrieve the URLs of the 100 top-ranked web pages, whose content we then retrieve via *wget*[1]. Trying to restrict the query results to those web pages that actually address the music band under consideration, we add domain-specific keywords to the query, which yields the following four query schemes:

- *"B"+music* (abbreviated as M in the following)
- *"B"+music+review* (MR)
- *"B"+music+members* (MM)
- *"B"+lineup+music* (LUM)

Discarding all markup tags, we eventually obtain a plain text representation of each web page.

Named Entity Detection

There is a large amount of literature on the topic of named entity detection. A good introduction can be found, for example, in [1]. For this work, we follow a quite simple approach. First, we extract all 2-, 3-, and 4-grams from the plain text representation of the web pages.[2] Subsequently, some basic filtering is performed. We exclude those N-grams whose substrings contain only one character and retain only those N-grams whose tokens all have their first letter in upper case and all remaining letters in lower case. Finally, we use the *iSpell English Word Lists*[3] to filter out those N-grams which contain at least one substring that is a common speech word. The remaining N-grams are regarded as potential band members.

Rule-based Linguistic Analysis

Having determined the potential band members, we perform a simple linguistic analysis to obtain the actual instrument of each member. Similar to the approach proposed in [3] for finding hyponyms in large text corpora, we define the following rules and apply them on the potential band members.

1. *M* plays the *I*
2. *M* who plays the *I*
3. *R M*
4. *M* is the *R*
5. *M*, the *R*
6. *M* (*I*)
7. *M* (*R*)

In these rules, *M* is the potential band member, *I* is the instrument, and *R* is the role *M* plays within the band (singer, guitarist, bassist, drummer, keyboardist). For *I* and *R*, we use synonym lists to cope with the use of multiple terms for the same concept (e.g. *percussion* and *drums*). We further count on how many of the web pages each rule applies for each M and I (or R).

Rule Selection

These counts are document frequencies (DF) since they indicate, for example, that on 24 web pages *Ralf Scheepers* is said to be the singer of the band *Primal Fear* according to rule 6 (on 6 pages according to rule 3, and so on). To reduce uncertain information, we filter out those rules whose DF is below a threshold expressed as a fraction of the DF of the highest scored rule (according to the DF score of all applying rules for the band under consideration).[4] Finally, for every instrument, the rule with

[1] *http://www.gnu.org/software/wget*

[2] We assume no artist name to comprise more than four single names.

[3] *http://wordlist.sourceforge.net*

[4] In our experiments we used 0.2 of the maximum DF as threshold.

the highest DF is selected and the respective (member, instrument)-pair is predicted.

3 EVALUATION AND RESULTS

To evaluate our approach, we compiled a ground truth based on one author's private music collection. Since this is a quite labor-intensive task, we restricted the collection to 51 bands, with a strong focus on the genre *Metal*. The chosen bands vary strongly with respect to their popularity (some are very well known, like *Metallica*, but most are largely unknown, like *Powergod*, *Pink Cream 69*, or *Regicide*). We gathered the current line-up of the bands by consulting *Wikipedia*[5], *allmusic*[6], *Discogs*[7], or the band's web site. Finally, our ground truth contained 240 members with their respective instruments.

We use three different string comparison methods to evaluate our approach. First, we perform *exact string matching*. Addressing the problem of different spelling for the same artist (e.g. the drummer of *Tiamat*, *Lars Sköld*, is often referred to as *Lars Skold*), we also evaluate the approach on the basis of a *canonical representation* of each band member. To this end, we perfom a mapping of similar characters to their stem, e.g. *ä, à, á, å, æ* to *a*. Furthermore, to cope with the fact that many artists use nicknames or abbreviations of their real names, we apply an *approximate string matching* method. According to [2], the so-called *Jaro-Winkler similarity* is well suited for personal first and last names since it favors strings that match from the beginning for a fixed prefix length (e.g. *Edu Falaschi* vs. *Eduardo Falaschi*, singer of the Brazilian band *Angra*). We use a *level two distance function* based on the Jaro-Winkler distance metric, i.e. the two strings to compare are broken into substrings (first and last names, in our case) and the similarity is calculated as the combined similarities between each pair of tokens. We assume that the two strings are equal if their Jaro-Winkler similarity is above 0.9. For calculating the distance, we use the open-source Java toolkit *SecondString*[8].

Table 1 shows the overall recall of the (band member, instrument)-pairs on the ground truth. A (member, instrument)-pair is only considered as correct if both the member and the instrument are predicted correctly. As for the influence of the query scheme, no significant difference could be made out between the *M*, the *MR*, and the *MM* settings. In contrast, the *LUM* scheme performed much worse, so we will exclude it in future experiments.

To estimate the goodness of the results given in Table 1, we analyzed, for the best performing query scheme *M*, how many of the actual band members (according to the ground truth) occur at least once in the retrieved web pages, i.e. for every band B, we calculated the recall, on the ground truth, of the N-grams extracted from B's web pages. We verified that no band members were erroneously discarded in the N-gram selection process. Over all band members, we obtained recall values of 56.00%, 57.64%, and 63.44% using exact matching, similar character mapping, and Jaro-Winkler distance, respectively. Taking these upper limits into account, the recall values given in Table 1 are quite promising for this preliminary study.

Table 1. Recall, in percent, of the (member, instrument)-pairs on the ground truth for different query schemes and string distance functions.

	exact	*similar char*	*L2-JaroWinkler*
M	34.76	37.14	39.05
MR	34.11	36.45	37.85
MM	35.98	36.92	37.38
LUM	26.64	27.57	27.57

Taking a qualitative look on the results, good performance was achieved for those bands whose principal members spent a long time in the band and are still members, regardless of the popularity of the band. For example, all (member, instrument)-pairs were correctly identified for the very famous *Iron Maiden*, but also for the less known *Edguy* and *Pink Cream 69*. On the other hand, our approach obviously has problems with heavy band member fluctuations, especially if a very famous member left the band after years of participation. A good example of this is *Nightwish*, whose long-term singer *Tarja Turunen* left the band in 2006. Moreover, since we restricted instrument detection to the five most popular ones used in Rock bands, the approach cannot deal with bands like *Apocalyptica*, comprising three cellists and one drummer.

4 FUTURE WORK

As for future work, we will try to improve performance by using more sophisticated rules and named entity detection approaches. Furthermore, we aim at deriving complete band histories (by searching for dates when a particular artist joined or left a band). This would allow us to create time-dependent relationship networks that could be used to derive a similarity measure. One possible application for this research is the creation of a domain-specific search engine for music artists, which is our ultimate aim.

5 ACKNOWLEDGEMENTS

This research is supported by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung (FWF) under project number L112-N04 and by the Vienna Science and Technology Fund (WWTF) under project number CI010 (Interfaces to Music).

6 REFERENCES

[1] J. Callan and T. Mitamura. Knowledge-Based Extraction of Named Entities. In *Proc. of the 11th Intl. Conf. on Information and Knowledge Management*, McLean, VA, USA, November 2002.

[2] W. W. Cohen, P. Ravikumar, and S. E. Fienberg. A Comparison of String Distance Metrics for Name-Matching Tasks. In *Proc. of the IJCAI-03 Workshop on Information Integration on the Web*, Acapulco, Mexico, August 2003.

[3] M. A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proc. of the 14th Conf. on Computational Linguistics - Vol. 2*, Nantes, France, August 1992.

[5] *http://www.wikipedia.org*

[6] *http://www.allmusic.com*

[7] *http://www.discogs.com*

[8] *http://secondstring.sourceforge.net*