Towards Semantic Music Information Extraction from the Web Using Rule Patterns and Supervised Learning

Peter Knees and Markus Schedl
Department of Computational Perception, Johannes Kepler University, Linz, Austria
peter.knees@jku.at, markus.schedl@jku.at

ABSTRACT

We present first steps towards automatic Music Information Extraction, i.e., methods to automatically extract semantic information and relations about musical entities from arbitrary textual sources. The corresponding approaches allow us to derive structured meta-data from unstructured or semi-structured sources and can be used to build advanced recommendation systems and browsing interfaces. In this paper, several approaches to identify and extract two specific semantic relations from related Web documents are presented and evaluated. The addressed relations are members of a music band ($band-members$) and artists' discographies ($artist - albums, EPs, singles$). In addition, the proposed methods are shown to be useful to relate (Web-)documents to musical artists. For all purposes, supervised learning approaches and rule-based methods are systematically evaluated on two different sets of Web documents.

Categories and Subject Descriptors

J.5 [**Arts and Humanities**]: *Music*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*

General Terms

Algorithms

Keywords

Music Information Extraction, Band-Member Relationship, Discography Extraction

1. MOTIVATION AND INTRODUCTION

Measuring similarity between artist, tracks or other musical entities — be it audio-based, Web-based, or a combination of both — is a key concept for music retrieval and recommendation. However, the type of relations between these entities, i.e., *what* makes them similar, is often neglected. Especially in the music domain, the number of potential relations between two entities is large. Such relations comprise, e.g., cover versions of songs, live versions, re-recordings, remixes, or mash-ups. Semantic high-level concepts such as "*song X* was inspired by *artist A*" or "*band B* is the new band of *artist A*" are very prominent in many users' conception and perception of music and should therefore be given attention in similarity estimation approaches. By focusing solely on acoustic properties, such relations are hard to detect (as can be seen, e.g., from research on cover version detection [7]).

A promising approach to deal with the limitations of signal-based methods is to exploit *contextual* information (for an overview see, e.g., [16]). Recent work in music information retrieval has shown that at least some cultural aspects can be modeled by analyzing extra-musical sources (often referred to as *community metadata* [25]). In the majority of work, this data — typically originating from Web sources and user data — is used for description/tagging of music (e.g., [10, 23, 24]) and assessment of similarity between artists (e.g., [17, 21, 22, 25]). However, while for these tasks standard information retrieval (IR) methods that reduce the obtained information to simple representations such as the bag-of-words model may suffice, important information on entities like artists' full names, band member names, album and track titles, related artists, as well as some music specific concepts like instrument names and musical styles may be dismissed. Addressing this issue, essential progress towards identifying relevant entities and, in particular, relations between these could be made. These kinds of information would also be highly valuable to automatically populate music-specific ontologies, such as the Music Ontology[1] [15].

In this paper, we aim at developing automatic methods to discover semantic relations between musical entities by analyzing texts from the Web. More precisely, to assess the feasibility of this goal, we focus on two specific sub-tasks, namely *automatic band member detection*, i.e., determining which persons a band consists (or consisted) of, and *automatic discography extraction*, i.e., recognition of released records (i.e., albums, EPs, and singles). Band member detection is strongly related to one of the central tasks of information extraction (IE) and named entity detection (NED), i.e., the recognition of persons' names in documents. While person's names typically exhibit some common patterns in terms of orthography and number of tokens, detection of artist names and band members is a bigger challenge as they frequently comprise or consist of nicknames, pseudonyms, or just a symbol (cf. *Prince* for a limited time). Discog-

WOMRAD 2011 2nd Workshop on Music Recommendation and Discovery, colocated with ACM RecSys 2011 (Chicago, US)

[1] `http://www.musicontology.com`

raphy detection in unstructured text is an even more challenging task as song or album names (release names in the following) are not bound to any conventions. That is, release names can consist of an unknown number of tokens (including zero tokens, cf. *The Beatles*'s "white album", or *Weezer*'s "blue", "green", and "red" albums, which might lead to inconsistent references on different sources), just special characters (e.g., *Justice*'s "Cross"), a differential equation (track 2 on *Aphex Twin*'s "Windowlicker" single), or whole paragraphs (e.g., the full title of a *Soulwax* album often abbreviated as *Most of the remixes* consists of 552 characters). Especially the last example demonstrates some of the challenges of a discography-targeted named entity recognition approach as the full album title itself exhibits linguistic structures and even contains another band's name (*Einstürzende Neubauten*). Hence, general methods not tailored to (or even aware of) music-related entities might not be able to deal with such specifics.

To investigate the potential and suitability of language-processing-based approaches for semantic music information extraction from (Web-)texts, two strategies commonly used in IE tasks are explored in this paper: manual tailoring of rule patterns to extract entities of interest (the "knowledge engineer" approach) and automatic learning of patterns from labeled data (supervised learning). Since particularly for the latter, pre-labeled data is required — which is difficult to obtain for most types of semantic relations — band-membership and discography extraction are, from our point of view, good starting points as these types of information are also largely available in a structured format (e.g., via Web services such as MusicBrainz[2]). In addition, the methods presented are also applied to relate documents to musical artists, which is useful for further tasks such as automatic music-focused crawling and indexing of the Web. In the bigger picture, these are supposed to be but the first steps towards a collection of methods to identify high-level musical relations between pieces, like cover versions, variations, remasterings, live interpretations, medleys, remixes, samples, etc. As some of these concepts are (partly) deducible from the audio signal itself, well considered methods for combining information from the audio with (Web-based) meta-information are required to automatically discover such relations.

2. RELATED WORK

The two music information extraction tasks addressed in this paper, i.e., band member and discography extraction, are specific cases of relation extraction. Since in the scenarios considered in this paper, one of the relational concepts is considered to be known (i.e., the band a text deals with), semantic relation extraction is reduced to named entity recognition and extraction tasks (i.e., extraction of band members and released records). Named entity recognition itself is a well-researched topic (for an overview see, e.g., [4]) and comprises the identification of proper names in structured or unstructured text as well as the classification of these names by means of rule-based or supervised learning approaches. While rule-based methods rely on experts that uncover patterns for the specific task and domain, supervised learning approaches require large amounts of labeled training data (which could, for instance, also stem from an ontology (cf. [1]). For the music domain – despite the numerous contributions that exploit Web-based sources to describe music or to derive similarity (cf. Section 1) – the number of publications aiming at extracting factual meta-data for musical entities by applying language processing methods is rather small.

In [19], we propose a first step to automatically extract the line-up of a music band, i.e., not only the members of a band but also their corresponding instruments and roles. As data source up to 100 Web documents for each band B, obtained via Google queries such as *"B" music*, *"B" music members*, or *"B" lineup music*, are utilized. From the retrieved pages, n-grams (where $n = \{2, 3, 4\}$), whose tokens consist of capitalized, non-common speech words of length greater than one are extracted. For band member and role extraction, a Hearst pattern approach (cf. [9]) is applied to the extracted n-grams and their surrounding text. The seven patterns used are 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), and 7. M (R), where M is the n-gram/potential band member, I an instrument, and R a role. For I and R, roles in a "standard rock band line-up", i.e., singer, guitarist, bassist, drummer, and keyboardist, as well as synonyms of these, are considered. After extraction, the document frequency of each rule is counted, i.e., on how many Web pages each of the above rules applies. Entities that occur on a percentage of band B's Web pages that is below a given threshold are discarded. The remaining member-role relations are predicted for B. In this paper, evaluation of the presented approaches is also carried out on the best-performing document set from [19] and compared against the Hearst pattern approach.

In [18], we investigate several approaches to determine the country of origin for a given artist, including an approach that performs keyword spotting for terms such as "born" or "founded" in the context of countries' names on Web pages. Another approach for country of origin determination is presented in [8]. Govaerts and Duval use selected Web sites and services, such as Freebase[3], Wikipedia[4], and Last.fm[5]. Govaerts and Duval propose three heuristics to determine the artist's country of origin using the occurrences of country names in biographies (highest overall occurrence, strongly favoring early occurrences, weakly favoring early occurrences). In [6], Geleijnse and Korst apply patterns like G *bands such as* A, *for example* A_1 *and* A_2, or M *mood by* A (where G represents a genre, A an artist name, and M a possible mood) to unveil genre-artist, artist-artist, and mood-artist relations, respectively.

While these music-specific information extraction methods mainly build upon few simple patterns or term frequency statistics, the work presented in this paper aims at incorporating more general methods that take advantage of linguistic features of the underlying texts and automatically learn models to derive musical entities annotated examples.

3. METHODOLOGY

The methods presented in this paper make use of the linguistic properties of texts related to music bands. To assess this information, for both approaches investigated (rule-based and supervised-learning-based), several pre-processing

[2] http://musicbrainz.org/

[3] http://www.freebase.com

[4] http://www.wikipedia.org

[5] http://last.fm

steps are required to obtain these linguistic features. Apart from initial preparation steps such as markup removal (if necessary), text tokenization (i.e., splitting the text into single tokens based on white spaces) and sentence splitting (based on punctuation), this comprises the following steps:

1. **Part-of-Speech Tagging (PoS)**: assigns PoS tags to tokens, i.e., annotates each token with its linguistic category (noun, verb, preposition, etc.), cf. [3].

2. **Gazetteer Annotation**: annotates occurrences of pre-defined keywords known to represent a specific concept, e.g., company names or persons' (first) names. These annotations can be used as look-up information for subsequent steps (see below). For the music domain, in this step, we also include lists of musical genres, instruments, and band roles, as well as a list of country names, cf. [11].

3. **Transducing Step**: identifies named entities such as persons, companies, locations, or dates using manually generated grammar rules. These rules can include lexical expressions, PoS information, look-up entities extracted via the gazetteer, or any other type of available annotation.

For all of these steps the functionalities included in the GATE software package (General Architecture for Text Engineering [5]) are utilized. In GATE's transducing step, detection of the different kinds of named entities is performed simultaneously in an interwoven process, i.e., decisions whether proper names represent persons or organizations are made after a number of shared intermediate steps. For instance, for person detection, information on first names and titles obtained from the gazetteer annotations are combined with information on initials, first names, surnames, and endings detected from orthographic characteristics (e.g., capitalization) and PoS tags. Finally, persons' surnames are removed if they contain certain stopwords or can be attributed to an organization. Details about this process can be found in Appendix F of the GATE User Guide[6].

The transducing step is also where we add additional rule-patterns designed to detect band members, releases, and artist names as described in the following section.

3.1 Rule-Pattern Approach

The first approach to extract music-related entities consists of generating specific rules that operate on the annotations obtained in the pre-processing steps. This requires the labor-intense task of manually detecting textual patterns that indicate certain entities in exemplary documents and writing (generalized) rules suited to capture other entities of the same concept also in new documents. For this purpose, for a set of 83 artists/bands, related Web pages such as band profiles and biographies from Last.fm, Wikipedia, and allmusic[7] are examined. Based on the made observations, rules that consider orthographic features, punctuation, surrounding entities (such as those identified via the gazetteer lists), and surrounding keywords are designed. The rules are formalized as so-called *JAPE grammars*[8] that are used in the transducer step of GATE. The complete set of JAPE grammars for music-specific entity recognition can be found in Appendix B of [11] and can also be obtained by contacting the authors. In the following, we show one exemplary (and easily accessible) rule for each concept to demonstrate idea and structure behind the rule-patterns for band member, media, and artist name extraction, respectively.

For the purpose of band member extraction, a JAPE grammar rule that aims at finding band members by searching for information about members leaving or joining the band is given as:

```
Rule : leftJoinedBand (
( ( MemberName ) ) : BandMember
({Token.string == "had"} | {Token.string == "has"})?
({Token.string == "left"} |
 {Token.string == "joined"} |
 {Token.string == "rejoined"} |
 {Token.string == "replaced"})
)--> :BandMember.Member =
     {kind = "BandMember", rule = "leftJoinedBand"}
```

To extract record releases, the following rule matches patterns that start with the potential media name (optionally in quotation marks) and point to production, release, performance, or similar events in the past or future:

```
Rule : MediaPassivReleased (({Token.string == "\""})?
( ( Medium ) ):Media
({Token.string == "\""})?
({Token.string == "was"} |
 ({Token.string == "will"} {Token.string == "be"}))
({Token.string == "released"} |
 {Token.string == "issued"} |
 {Token.string == "produced"} |
 {Token.string == "recorded"} |
 {Token.string == "played"} |
 {Token.string == "performed"} ))--> :Media.Media =
     {kind = "Media", rule = "MediaPassivReleased"}
```

To identify occurrences of band names, the following rule focuses on the entity occurring before terms such as *was founded* or *were supported*:

```
Rule : Formed (
( ( BandN ) ) : BandName({Token.string == "was"} |
{Token.string == "were"})
({Token.string == "formed"} |
 {Token.string == "supported"} |
 {Token.string == "founded"}))--> :BandName.bandname =
     {kind = "Band", rule = "Formed"}
```

Elaborating such rules is a tedious task and (especially in heterogeneous data environments such as the Web) unlikely to generalize well and cover all cases. Therefore, in the next section we describe a supervised learning approach that makes use of automatically labeled data.

3.2 Supervised Learning Approach

Instead of manually examining unstructured text for occurrences of musical entities and potential patterns to identify them, the idea of this approach is to apply a supervised learning algorithm to a set of pre-annotated examples. Using the learned model, relevant information should then be found also in new documents. Several approaches, more precisely several types of machine learning algorithms, have been proposed for automatic information extraction tasks, such as hidden-markov-models [2], decision trees [20], or support vector machines (SVM) [12]. Since the latter demonstrates that SVMs may yield results that rival those of optimized rule-based approaches, SVMs are chosen as classifier for the tasks at hand (for more details see [12, 13])

[6] http://gate.ac.uk/userguide/

[7] http://www.allmusic.com

[8] Acronym for Java Annotation Patterns Engine

For training of the SVMs, a set of documents that contain annotations of the entities of interest is required. Since also this step can be labor intense, we opted for an automatic annotation approach. For the collection of training documents, ground truth information (on band member history and band discography) is obtained by either manually compiling lists or by invoking Web services such as MusicBrainz or Freebase. Using this information, occurrences of the band name, its members (full name as well as last name only), and releases are annotated using regular expressions.

Construction of the features and SVM training is carried out as described by Li et al. [12]. First, for each token, a feature vector representation has to be obtained. In the given scenario, for each token, its content (i.e., the actual string), orthographic properties, PoS information, gazetteer-based entity information, and identified person entities are considered. In a second scenario, in addition to these, also the output of the rule-based approach (more precisely, the name of the rule responsible for prediction of an entity) serves as an input feature. Ideally, this incorporates indicators of high relevance and allows for supervised selection of the manually generated rules for the final predictions. For each prediction task, the corresponding annotation type is also added to the features as target class.

To construct the feature vectors, the training corpus is scanned for all occurring values of any of the considered attributes (i.e., annotations). Then, each token is represented by a vector where each distinct annotation value corresponds to one dimension which is set to 1 if the token is annotated with the corresponding value. In addition, the context of each token (consisting of a window that includes the 5 preceding and the 5 subsequent tokens) is incorporated. This is achieved by creating an SVM input vector for each token that is a concatenation of the feature vectors of all tokens in the context window. To reflect the distance of the surrounding tokens to the actual token (i.e., the center of the window), a reciprocal weighting is applied, meaning that "the nonzero components of the feature vector corresponding to the j^{th} right or left neighboring word are set to be equal to $1/j$ in the combined input vector." [12]. In our experiments, this typically results in feature vectors with approximately 1.5 million dimensions.

In the SVM learning phase, the input vectors corresponding to every single token in all training documents serve as examples. According to the central idea of [12], two distinct SVM classifiers are trained for each concept of interest. The first classifier is trained to predict the beginning of an entity (i.e., to classify whether a token is the first token of an entity), the second to predict the end (i.e., whether a token is the last token of an entity). To deal with the unbalanced distribution of positive and negative training examples, a special form of SVMs is used, namely an SVM with uneven margins [14]. From the obtained predictions of start and end positions, actual entities, as well as corresponding confidence scores, are determined in a post-processing step. First, start tokens without matching end token, as well as end tokens without matching start token are removed. Second, entities with a length (in terms of the number of tokens) that does not match any training example's length are discarded. Third, a confidence score is calculated based on a probabilistic interpretation of the SVM output for all possible classes. More precisely, for each entity, the conjunction of the Sigmoid transformed SVM output probabilities of start and end token is calculated for each possible output class. Finally, the class (label) with the highest probability is predicted for the entity if its probability is greater than 0.25. The probability of the predicted class serves as a confidence score.

3.3 Entity Consolidation and Prediction

From the extraction step (either rule- or learning-based), for each processed text and each concept of interest, a list of potential entities is obtained. For each band, the lists from all texts associated with the band are joined and the occurrences of each entity as well as the number of texts an entity occurs in are counted (term and document frequency, respectively). The joined list usually contains a lot of noise and redundant data, calling for a filtering and merging step. First, all entities extracted by the learning-based method that have a confidence score below 0.5 are removed since they are more likely to not represent band members than representing band members according to the classification step. On the cleaned list, the same observations as described in [19] can be made. For instance, on the list of extracted band members, some members are referenced with different spellings (*Paavo Lötjönen* vs. *Paavo Lotjonen*), with abbreviated first names (*Phil Anselmo* vs. *Philip Anselmo*), with nicknames (*Darrell Lance Abbott* vs. *Dimebag Darrell* or just *Dimebag*), or only by their last name (*Iommi*). On the discography lists, release names are often followed by additional information such as release year or type of release. This is dealt with by introducing an approximate string matching function, namely the level-two Jaro-Winkler similarity, cf. [19].[9] For both entity types, this type of similarity function is suited well as it assigns higher matching scores to pairs of strings that start with the same sequence of characters. In the level-two variant, the two entities to compare are split into substrings and similarity is calculated as an aggregated similarity of pairwise comparison of the substrings. To reduce redundancies, two entities are considered synonymous and thus merged if their level-two Jaro-Winkler similarity is above 0.9. In addition, to deal with the occurrence of last names, an entity consisting of one token is considered a synonym of another entity if it matches the other entity's last token.

This consolidated list is usually still noisy, calling for additional filtering steps. To this end, two threshold parameters are introduced. The first threshold, $t_f \in \mathbb{N}^0$, determines the minimum number of occurrences of an entity (or its synonyms) in the band's set to get predicted. The second threshold, $t_{df} \in [0...1]$ controls the lower bound of the fraction of texts/documents associated with the band an entity has to occur in (document frequency in relation to the total number of documents per band). The impact of these two parameters is systematically evaluated in the following section.

4. EVALUATION

To assess the potential of the proposed approaches and to measure the impact of the parameters, systematic experiments are conducted. This section details the used test collections as well as the applied evaluation measures and reports on the results of the experiments.

[9] For calculation, the open-source Java toolkit *SecondString* (`http://secondstring.sourceforge.net`) is utilized.

4.1 Test Collections

For evaluation, two collections with different characteristics are used – the first a previously published collection used in [19], the second a larger scale test collection consisting of band biographies.

4.1.1 Metal Page Sets

The first collection is a set of Web pages introduced in [19]. This set consist of Google's 100 top-ranked Web pages retrieved using the query *"band name"music members* (cf. Section 2) for 51 Rock and Metal bands (resulting in a total of 5,028 Web pages). In [19], this query setting yielded best results and is therefore chosen as reference for the task of band-member extraction. As ground truth, the membership-relations that include former members are chosen (i.e., the M_f ground truth set of [19]). For this evaluation collection also the results obtained by applying the Hearst patterns proposed in [19] are available, allowing for a direct comparison of the approaches' band member extraction capabilities.

For the discography extraction evaluation, no reference data is available in the original set. Therefore – and since the discography of the contained bands has changed since the creation of the set – a new Web crawl has been conducted to retrieve recent (and more related) data. Since the aim of this new set is to extract released media, for each of the 51 bands in the metal set the query *"band name" discography* is sent to Google and the top 100 pages are downloaded (resulting in a total of 5,090 Web pages). To obtain a discography ground truth, titles of albums, EPs, and singles released by each band are downloaded from MusicBrainz.

To speed up processing of the collections, all Web pages with a file size over 100 kilobyte are discarded resulting in set sizes of 4,561 and 4,625 documents for the member set and the discography set, respectively. Evaluation of the supervised learning approach is performed as a 2-fold cross validation (by splitting the band set and separating the associated Web pages), where in each fold a random sample of 100 documents is drawn for training.

4.1.2 Biography Set

The second test collection is a larger scale collection consisting only of band biographies to be found on the Web. Biographies are investigated as they should contain both information on (past) band members and information on (important) released records.

Starting from a snapshot of the MusicBrainz database from December 2010, all artists marked as bands and all corresponding band members as well as albums, EPs, and singles are extracted. In addition, also band-membership information from Freebase[10] is retrieved and merged with the MusicBrainz information to make the ground truth data set more comprehensive. After this step, band-membership information is available for 34,238 bands. For each band name, the echonest API[11] is invoked to obtain related biographies. Using the echonest's Web service, related biographies (e.g., from Wikipedia, Last.fm, allmusic, or Aol Music[12]) can be conveniently retrieved in plain text format. Since among the provided biographies for a band, duplicates or near-duplicates, as well as only short snippets can be observed, (near-)duplicates as well as biographies consisting of less than 100 characters are filtered out. After filtering (near-)duplicates and snippets, for 23,386 bands (68%) at least one biography remains. In total, a set of 38,753 biographies is obtained. To keep processing times short, furthermore all documents that contain more than 10 megabyte of annotations after the initial processing step are filtered out.

For training of the supervised learner, a random subset of 100 biographies is chosen. All biographies by any artist that is part of the training set are removed from the test set, resulting in a final test set of 37,664 biographies by 23,030 distinct bands.

In comparison to the first test sets, i.e., the Metal page sets, the biography set contains more bands, more specific documents in a homogeneous format (i.e., biographies instead of semi-structured Web pages from various sources), but less associated documents (in average 1.63 documents per band, as opposed to an average of 90 documents per band for the Metal page set).

4.2 Evaluation Metrics

For evaluation, *precision* and *recall* are calculated separately for each band and averaged over all bands to obtain a final score. The metrics are defined as follows:

$$precision = \begin{cases} \frac{|T \cap P|}{|P|} & \text{if } |P| > 0 \\ 1 & \text{otherwise} \end{cases} \quad (1)$$

$$recall = \frac{|T \cap P|}{|T|} \quad (2)$$

where P is the set of predicted entities and T the ground truth set of the band. To assess whether an extracted entity is correct, again the level-two Jaro-Winkler similarity (see Section 3.3) is applied. More precisely, if the Jaro-Winkler similarity between a predicted entity and an entity contained in the ground truth is greater than 0.9, the prediction is considered to be correct. Furthermore, if a predicted band member name consist of only one token, it is considered correct, if it matches with the last token of a member in the ground truth. These weakened definitions of matching allow for tolerating small spelling variations, name abbreviations, extracted last names, additional information of releases, as well as string encoding differences.

For comparison with the Hearst pattern approach for band member detection on the Metal page set, it has to be noted that in [19], calculation of precision and recall is done on the full set of bands and members (and their corresponding roles), yielding global precision and recall values, whereas here, the evaluation metrics are calculated separately for each band and are then averaged over all bands to remove the influence of a band's size. Using the global evaluation scheme, e.g., orchestras are given far more importance than, for instance, duos in the overall evaluation, although for a duo, the individual members are generally more important than for an orchestra. Therefore, in the following, the different approaches are compared based on macro-averaged evaluation metrics (calculated using the arithmetic mean of the individual results).

4.3 Evaluation Results

In the following, the proposed rule-patterns, the SVM approach, as well as the SVM approach that utilizes the out-

[10] `http://www.freebase.com`

[11] `http://developer.echonest.com`

[12] `http://music.aol.com`

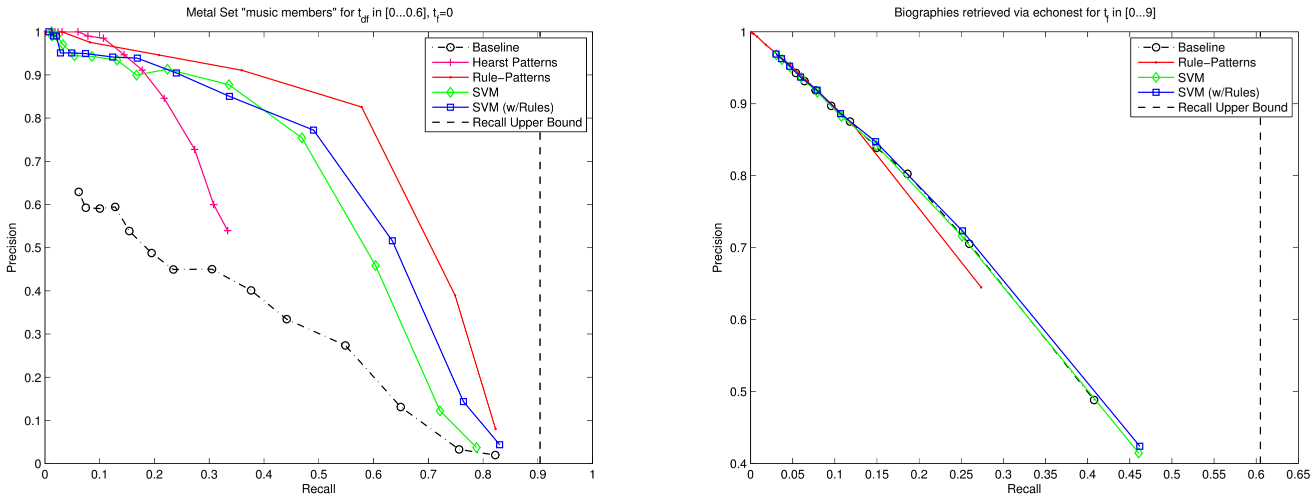

Figure 1: Precision-recall plots for band-member prediction on the Metal page set (left) and on the biography set (right). Curves are obtained by systematically varying threshold parameters (t_{df} and t_f for Metal page set and biography set, respectively). Precision and recall values macro-averaged over all bands in the corresponding test set.

put of the rule-patterns are compared for the tasks of band-member detection and discography extraction. For detecting band-members, a baseline reference consisting of the person entity prediction functionality of GATE is provided. On the Metal page set, band-member prediction is further compared to the Hearst pattern approach from [19]. For the task of discography extraction, no such reference is available. For all evaluations, an additional upper bound for the recall is calculated. This upper bound is implied by the underlying documents, since band members and releases that do not occur on any of the documents can not be predicted.

4.3.1 Band-Member Detection

The left part of Figure 1 shows precision-recall curves for the different band member detection approaches on the Metal page set. For a systematic comparison with the Hearst pattern approach, the t_{df}, i.e., the threshold that determines on which fraction of a band's total documents a band member has to appear on to be predicted, is varied. It can be seen that the rule-based approach clearly performs best. Also SVM and SVM using the rules output outperform the Hearst pattern approach. It becomes apparent that on the Metal set, rule patterns, the GATE person baseline, and the supervised approaches can yield recall values close to the upper bound, i.e., these approaches capture nearly all members contained in the documents at least once. For the Hearst patterns, recall remains low. However, when comparing the Hearst patterns, it has to be noted that this approach was initially designed to also detect the roles of the band members — a feature none of the other approaches is capable of.

Since on the biography set only 1.63 documents per band are available on average, variation of the t_{df} threshold is not as interesting as on the Metal page set. Therefore, the right part of Figure 1 depicts curves of the proposed approaches with varied values of t_f, i.e., the threshold that determines how often an entity has to be detected to be predicted as a band member. On this set, the supervised learning approaches tend to outperform the rule-based extraction approach slightly. However, there is basically no difference between the SVM approaches and the baseline with the only exception that the SVM approaches can yield higher recall values. Another observation is that the upper recall boundary on the biography set is rather low at about 0.6.

4.3.2 Discography Extraction

For discography extraction the situation is similar as can be seen from Figure 2. Also for this task the rule-based approach outperforms the SVM approaches (this time also on the biography set). Recall is also close to the upper bound using SVMs on the Metal page set while on the biography set, none of the approaches is capable of reaching the already low upper recall boundary at 0.36. Conversely, on the biography set, all proposed approaches yield rather high precision values. However, due to the lack of a baseline reference, it is difficult to draw final conclusions about the quality of these approaches for the task of discography extraction.

What can be seen from both the evaluations on discography and band-member extraction is that – despite all work required – rule-patterns are preferable over supervised learning methods. Another consistent finding so far is that SVMs that utilize the output of the rule-pattern classification process are superior to SVMs without this information, but still inferior to the predictions of the rule-patterns alone.

The most unexpected result can be observed for band-member extraction on the biography set. None of the proposed methods outperforms the standard person detection approach by GATE. A possible explanation could be that the baseline itself is already high. Since biographies typically follow a certain writing style and consist — in contrast to arbitrary Web pages — mostly of grammatically well-formed sentences, natural language processing techniques such as PoS tagging perform better on this type of input. Thus, the person detection approach just works better on the biography data than on the Metal page set.

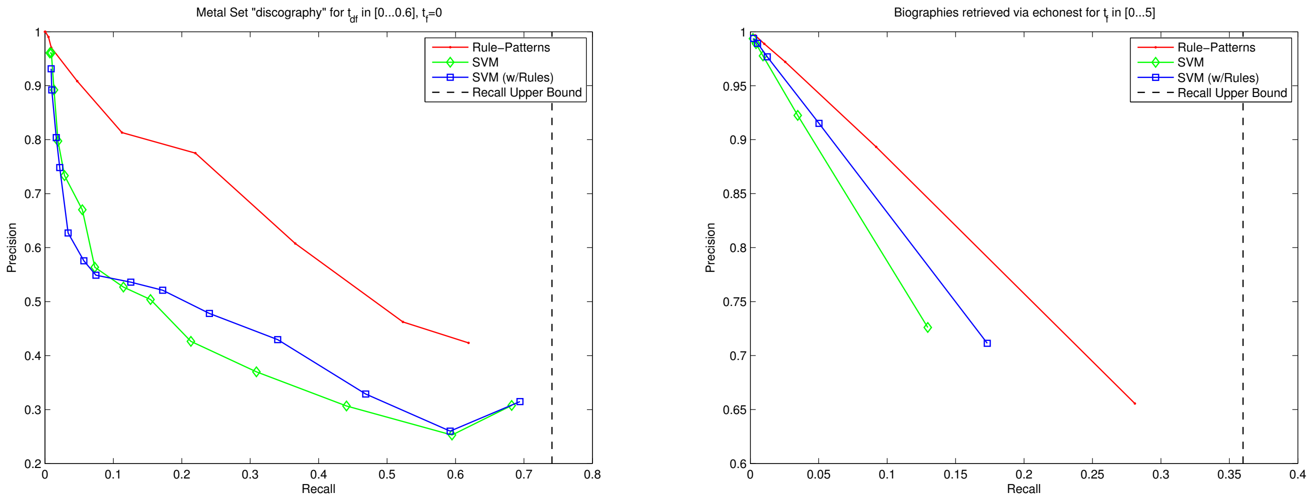

Figure 2: Precision-recall plots for discography detection on the Metal page set (left) and on the biography set (right). Settings as in Figure 1.

In terms of the different sources of data, i.e., the chosen test collections, it can be seen that using biographies, in general lower recall values (and higher precision values) should be expected. This can be seen also from the upper recall bounds that are rather low for both tasks. When using Web documents, more information can be accessed which results also in higher recall values. On the discography Metal set, a recall of 0.7 can be observed which is already close to the upper bound of 0.74. However, using Web documents requires considerations which documents to examine (e.g., by formulating an appropriate query to obtain many relevant pages) as well as dealing with a lot of noise in the data.

4.3.3 *Relating Documents to Artists*

In addition to the two main tasks of this paper, we also briefly investigate the applicability of the presented methods to identify the central artist or band in a text about music, which could be useful for future relation extraction tasks and tools such as music-focused Web crawling and indexing. To this end, we utilize the rule-patterns aiming at detecting occurrences of artists and train SVMs on occurrences of the name of the band a page belongs to. For prediction, the most frequently extracted entity with occurrences greater than a threshold t_f is selected. As a baseline, simple prediction of any sequence of capitalized tokens at the beginning of the text is chosen. The results can be seen in Figure 3. For this task, SVMs perform better than the rule-patterns. However, rather surprisingly, the highest recall value can be observed for the simple baseline.

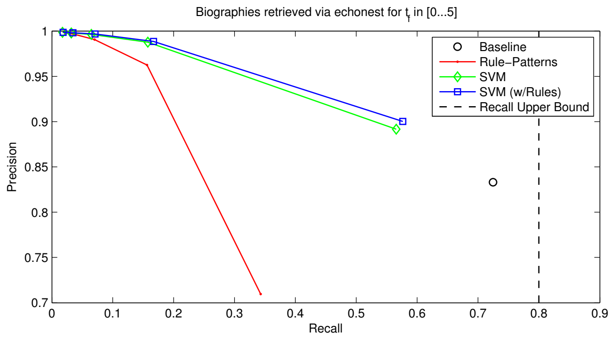

Figure 3: Precision-recall plots for discography detection on the biography set. Curves obtained by varying threshold parameter t_f. Precision and recall values averaged over all pages.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we presented first steps towards semantic Music Information Extraction. We focused on two specific tasks, namely determining the members of a music band and determining the discography of an artist (also explored on sets of bands). For both purposes, supervised learning approaches and rule-based methods were systematically evaluated on two different sets of documents. From the conducted evaluations, it became evident that manually generated rules yield superior results. Furthermore, it could be seen that careful selection of the underlying data source is crucial to achieve reliable results.

In general, the results obtained show great potential for these and also related tasks. By just focusing on biographies, even more highly relevant meta-information on music could be extracted. For instance, consider the following paragraph taken from the Wikipedia page of the *Alkaline Trio*:

"In September 2006, Patent Pending, the debut album by Matt Skiba's side project Heavens was released. The band consisted of Skiba on guitar and vocals, and Josiah Steinbrick (of hardcore punk outfit F-Minus) on bass. On the album, the duo were joined by The Mars Volta's Isaiah "Ikey" Owens on organ and Matthew Compton on drums and percussion."[13]

This short paragraph contains band-membership and line-up information for the *Alkaline Trio*, for the band *Heavens*, for the band *F-Minus*, and for the band *The Mars*

[13] `http://en.wikipedia.org/w/index.php?title=Alkaline_Trio&oldid=431587984`

Volta. In addition, discographical information for *Heavens*, genre information for *F-Minus*, and a nickname/alias for *Isaiah Owens* can be inferred from this small piece of text. Furthermore, relations between the mentioned bands ("side-project") as well as the mentioned persons (collaborations) can be discovered. Using further information extraction methods, in future work, it should be possible to capture at least some of this semantic information and relations and to advance the current state-of-the-art in music retrieval and recommendation. However, for systematic experimentation and targeted development, the creation of a comprehensive and thoroughly (manually) annotated text corpus for music seems unavoidable.

6. ACKNOWLEDGMENTS

Thanks are due to Andreas Krenmair for conceiving the music-related JAPE patterns and sharing his implementation. This research is supported by the Austrian Research Fund (FWF) under grants L511-N15 and P22856-N23.

7. REFERENCES

[1] H. Alani, S. Kim, D.E. Millard, M.J. Weal, W. Hall, P.H. Lewis, and N.R. Shadbolt. Automatic Ontology-Based Knowledge Extraction from Web Documents. *IEEE Intelligent Systems*, 18(1):14–21, 2003.

[2] D. M. Bikel, S. Miller, R. Schwartz, and R. Weischedel. Nymble: a High-Performance Learning Name-finder. In *Proc. 5th Conference on Applied Natural Language Processing*, 1997.

[3] E. Brill. A Simple Rule-Based Part of Speech Tagger. In *Proc. 3rd Conference on Applied Natural Language Processing*, 1992.

[4] J. Callan and T. Mitamura. Knowledge-Based Extraction of Named Entities. In *Proc. 11th International Conference on Information and Knowledge Management (CIKM)*, 2002.

[5] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proc. 40th Anniversary Meeting of the Association for Computational Linguistics*, 2002.

[6] G. Geleijnse and J. Korst. Web-based artist categorization. In *Proc. 7th International Conference on Music Information Retrieval (ISMIR)*, 2006.

[7] E. Gómez and P. Herrera. The song remains the same: Identifying versions of the same piece using tonal descriptors. In *Proc. 7th International Conference on Music Information Retrieval (ISMIR)*, 2006.

[8] S. Govaerts and E. Duval. A Web-Based Approach to Determine the Origin of an Artist. In *Proc. 10th International Society for Music Information Retrieval Conference (ISMIR)*, 2009.

[9] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proc. 14th Conference on Computational Linguistics - Vol. 2*, 1992.

[10] P. Knees. *Text-Based Description of Music for Indexing, Retrieval, and Browsing*. PhD thesis, Johannes Kepler Universität, Linz, Austria, 2010.

[11] A. Krenmair. Musikspezifische Informationsextraktion aus Webdokumenten. Diplomarbeit, Johannes Kepler Universität, Linz, Austria, 2010.

[12] Y. Li, K. Bontcheva, and H. Cunningham. SVM Based Learning System for Information Extraction. In J. Winkler, M. Niranjan, and N. Lawrence, eds., *Deterministic and Statistical Methods in Machine Learning*, vol. 3635 of *LNCS*. Springer, 2005.

[13] Y. Li, K. Bontcheva, and H. Cunningham. Adapting SVM for Data Sparseness and Imbalance: A Case Study on Information Extraction. *Natural Language Engineering*, 15(2):241–271, 2009.

[14] Y. Li and J. Shawe-Taylor. The SVM with uneven margins and Chinese document categorization. In *Proc. 17th Pacific Asia Conference on Language, Information and Computation (PACLIC)*, 2003.

[15] Y. Raimond, S. Abdallah, M. Sandler, and F. Giasson. The Music Ontology. In *Proc. 8th International Conference on Music Information Retrieval (ISMIR)*, 2007.

[16] M. Schedl and P. Knees. Context-based Music Similarity Estimation. In *Proc. 3rd International Workshop on Learning the Semantics of Audio Signals (LSAS)*, 2009.

[17] M. Schedl, P. Knees, and G. Widmer. A Web-Based Approach to Assessing Artist Similarity using Co-Occurrences. In *Proc. 4th International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2005.

[18] M. Schedl, C. Schiketanz, and K. Seyerlehner. Country of Origin Determination via Web Mining Techniques. In *Proc. IEEE International Conference on Multimedia and Expo (ICME): 2nd International Workshop on Advances in Music Information Research (AdMIRe)*, 2010.

[19] M. Schedl and G. Widmer. Automatically Detecting Members and Instrumentation of Music Bands via Web Content Mining. In *Proc. 5th Workshop on Adaptive Multimedia Retrieval (AMR)*, 2007.

[20] S. Sekine. NYU: Description of the Japanese NE system used for MET-2. In *Proc. 7th Message Understanding Conference (MUC-7)*, 1998.

[21] Y. Shavitt and U. Weinsberg. Songs Clustering Using Peer-to-Peer Co-occurrences. In *Proc. IEEE International Symposium on Multimedia (ISM): International Workshop on Advances in Music Information Research (AdMIRe)*, 2009.

[22] M. Slaney and W. White. Similarity Based on Rating Data. In *Proc. 8th International Conference on Music Information Retrieval (ISMIR)*, 2007.

[23] M. Sordo, C. Laurier, and O. Celma. Annotating Music Collections: How Content-based Similarity Helps to Propagate Labels. In *Proc. 8th International Conference on Music Information Retrieval (ISMIR)*, 2007.

[24] D. Turnbull, L. Barrington, and G. Lanckriet. Five Approaches to Collecting Tags for Music. In *Proc. 9th International Conference on Music Information Retrieval (ISMIR)*, 2008.

[25] B. Whitman and S. Lawrence. Inferring Descriptions and Similarity for Music from Community Metadata. In *Proc. International Computer Music Conference (ICMC)*, 2002.