Automatically Detecting Members and Instrumentation of Music Bands Via Web Content Mining

Markus Schedl[1] and Gerhard Widmer[1,2]

[1] Department of Computational Perception
Johannes Kepler University
Linz, Austria
markus.schedl@jku.at
http://www.cp.jku.at
[2] Austrian Research Institute for Artificial Intelligence
Vienna, Austria
gerhard.widmer@jku.at
http://www.ofai.at

Abstract. In this paper, we present an approach to automatically detecting music band members and instrumentation using web content mining techniques. To this end, we combine a named entity detection method with a rule-based linguistic text analysis approach extended by a rule filtering step. We report on the results of different evaluation experiments carried out on two test collections of bands covering a wide range of popularities. The performance of the proposed approach is evaluated using precision and recall measures. We further investigate the influence of different query schemes for the web page retrieval, of a critical parameter used in the rule filtering step, and of different string matching functions which are applied to deal with inconsistent spelling of band members.

1 Introduction and Context

Automatically retrieving textual information about music artists is a key question in text-based music information retrieval (MIR), which is a subfield of multimedia information retrieval. Such information can be used, for example, to enrich music information systems or music players [14], for automatic biography generation [1], to enhance user interfaces for browsing music collections [9,6,11,16], or to define similarity measures between artists, a key concept in MIR. Similarity measures enable, for example, creating relationship networks [3,13] or recommending unknown artists based on the favorite artists of the user (recommender systems) [17] or based on arbitrary textual descriptions of the artist or music (music search engines) [8].

Here, we present an approach that was developed for – but is not restricted to – the task of finding the members of a given music band and the respective instruments they play. In this 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). Since our approach relies

N. Boujemaa, M. Detyniecki, and A. Nürnberger (Eds.): AMR 2007, LNCS 4918, pp. 122–133, 2008.

on information provided on the web by various companies, communities, and interest groups (e.g. record labels, online stores, music information systems, listeners of certain music genres), it adapts to changes as soon as new or modified web pages incorporating the changes become available. Deriving (member, instrument)-assignments from web pages is an important step towards building a music information system whose database is automatically populated by reliable information found on the web, which is our ultimate aim.

The approach presented in this paper relates to the task of *named entity detection (NED)*. A good outline of the evolution of NED can be found in [2]. Moreover, [2] presents a knowledge-based approach to learning rules for NED in structured documents like web pages. To this end, document-specific extraction rules are generated and validated using a database of known entity names. In [10], information about named entities and non-named entity terms are used to improve the quality of new event detection, i.e. the task of automatically detecting, whether a given story is novel or not. The authors of [15] use information about named entities to automatically extract facts and concepts from the web. They employ methods including domain-specific rule learning, identifying subclasses, and extracting elements from lists of class instances.

The work presented in [4] strongly relates to our work as the authors of [4] propose a pattern-based approach to finding instances of concepts on web pages and classify them according to an ontology of concepts. To this end, the page counts returned by *Google* for search queries containing hypothesis phrases are used to assign instances to concepts. For the general geographic concepts (e.g. city, country, river) and well-known instances used in the experiments in [4], this method yielded quite promising results.

In contrast, the task which we address in this paper, i.e. assigning (member, instrument)-pairs to bands, is a more specific one. Preliminary experiments on using the page counts returned for patterns including instrument, member, and band names yielded very poor results. In fact, querying such patterns as exact phrases, the number of found web pages was very small, even for well-known bands and members. Using conjunctive queries instead did not work either as the results were, in this case, heavily distorted by famous band members frequently occurring on the web pages of other bands. For example, *James Hetfield*, singer and rhythm guitarist of the band *Metallica*, occurs in the context of many other Heavy Metal bands. Thus, he would likely be predicted as the singer (or guitarist) of a large number of bands other than *Metallica*. Furthermore, the page counts returned by *Google* are only very rough estimates of the actual number of web pages. For these reasons, we elaborated an approach that combines the power of *Google*'s page ranking algorithm [12] (to find the top-ranked web pages of the band under consideration) with the precision of a rule-based linguistic analysis method (to find band members and assign instruments to them).

The remainder of this paper is organized as follows. Section 2 presents details of the proposed approach. In Section 3, the test collection used for our experiments is introduced. Subsequently, the conducted experiments are presented

and the evaluation results are discussed in Section 4. Finally, Section 5 draws conclusions and points out directions for future research.

2 Methodology

The basic approach comprises four steps: *web retrieval*, *named entity detection*, *rule-based linguistic analysis*, and *rule selection*. Each of these are elaborated on in the following.

2.1 Web Retrieval

Given a band name B, we use *Google* to obtain 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* (abbreviated as MR in the following)
- *"B"+music+members* (abbreviated as MM in the following)
- *"B"+lineup+music* (abbreviated as LUM in the following)

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

2.2 Named Entity Detection

We employ a quite simple approach to NED, which basically relies on *detecting capitalization* and on *filtering*. First, we extract all *2-, 3-, and 4-grams* from the plain text representation of the web pages as we assume that the complete name of a band member comprises at least two and at most four single names, which holds for our test collection as well as for the vast majority of band members in arbitrary collections. 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*[2] 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.

2.3 Rule-Based Linguistic Analysis

Having determined the potential band members, we perform a linguistic analysis to obtain the actual instrument(s) of each member. Similar to the approach proposed in [7] for finding hyponyms in large text corpora, we define the following rules and apply them on the potential band members (and the surrounding text as necessary) found in the named entity detection step.

[1] *http://www.gnu.org/software/wget*
[2] *http://wordlist.sourceforge.net*

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).

2.4 Rule Selection According to Document Frequencies

These counts are document frequencies (DF) since they indicate, for example, that on 24 of the web pages returned for the search query "Primal Fear"+music *Ralf Scheepers* is said to be the singer of the band according to rule 6 (on 6 pages according to rule 3, and so on). The extracted information is stored as a set of quadruples (member, instrument, rule, DF) for every band. Subsequently, the DF given by the individual rules are summed up over all (member, instrument)-pairs of the band, which yields (member, instrument, $\sum$DF)-triples. To reduce uncertain membership predictions, we filter out the triples whose $\sum$DF values are below a threshold t_{DF}, both expressed as a fraction of the highest $\sum$DF value of the band under consideration. To give an example, this filtering would exclude, in a case where the top-ranked singer of a band achieves an accumulated rule DF ($\sum$DF) of 20, but no potential drummer scores more than 1, all potential drummers for any $t_{DF} > 0.05$. Thus, the filtering would discard information about drummers since they are uncertain for the band.

In preliminary experiments for this work, after having performed the filtering step, we predicted, for each instrument, the (member, instrument)-pair with the highest $\sum$DF value. Unfortunately, this method allows only for a $1 : m$ assignment between members and instruments. In general, however, an instrument can be played by more than one band member within the same band. To address this issue, for the experiments presented here, we follow the approach of predicting all (member, instrument)-pairs that remain after the filtering according to DF step described above. This enables an $m : n$ assignment between instruments and members.

3 Test Collection

To evaluate the proposed approach, we compiled a ground truth based on one author's private music collection. As this is a labor-intensive and time-consuming task, we restricted the dataset to 51 bands, with a strong focus on the genre *Metal*. The chosen bands vary strongly with respect to their popularity (some

Table 1. A list of all band names used in the experiments

Angra	Annihilator	Anthrax
Apocalyptica	Bad Religion	Black Sabbath
Blind Guardian	Borknagar	Cannibal Corpse
Century	Crematory	Deicide
Dimmu Borgir	Edguy	Entombed
Evanescence	Finntroll	Gamma Ray
Green Day	Guano Apes	Hammerfall
Heavenly	HIM	Iron Maiden
Iron Savior	Judas Priest	Krokus
Lacuna Coil	Lordi	Majesty
Manowar	Metal Church	Metallica
Motörhead	Nightwish	Nirvana
Offspring	Pantera	Paradise Lost
Pink Cream 69	Powergod	Primal Fear
Rage	Regicide	Scorpions
Sepultura	Soulfly	Stratovarius
Tiamat	Type O Negative	Within Temptation

are very well known, like *Metallica*, but most are largely unknown, like *Powergod*, *Pink Cream 69*, or *Regicide*). A complete list of all bands in the ground truth can be found in Table 1. We gathered the current line-up of the bands by consulting *Wikipedia*[3], *allmusic*[4], *Discogs*[5], or the band's web site. Finally, our ground truth contained 240 members with their respective instruments. We denote this dataset, that contains the current band members at the time we conducted the experiments (March 2007), as M_c in the following.

Since we further aimed at investigating the performance of our approach on the task of finding members that already left the band, we created a second ground truth dataset, denoted M_f in the following. This second dataset contains, in addition to the current line-up of the bands, also the former band members. Enriching the original dataset M_c with these former members (by consulting the same data sources as mentioned above), the number of members in M_f adds up to 499.

4 Evaluation

We performed different evaluations to assess the quality of the proposed approach. First, we calculated precision and recall of the predicted (member, instrument)-pairs on the ground truth using a fixed t_{DF} threshold. To get an impression of the goodness of the recall values, we also determined the upper bound for the recall achievable with the proposed method. Such an upper bound

[3] *http://www.wikipedia.org*
[4] *http://www.allmusic.com*
[5] *http://www.discogs.com*

exists since we can only find those members whose names actually occur in at least one web page retrieved for the artist under consideration. Subsequently, we investigate the influence of the parameter t_{DF} used in the rule filtering according to document frequencies. We performed all evaluations on both ground truth datasets M_c and M_f using each of the four query schemes.

We further employ 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 [5], 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 caluclated 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*[6].

4.1 Precision and Recall

We measured precision and recall of the predicted (member, instrument)-pairs on the ground truth. Such a (member, instrument)-pair is only considered correct if both the member and the instrument are predicted correctly. We used a threshold of $t_{DF} = 0.25$ for the filtering according to document frequencies (cf. Subsection 2.4) since according to preliminary experiments, this value seemed to represent a good trade-off between precision and recall.

Given the set of correct (band member, instrument)-assignments T according to the ground truth and the set of assignments predicted by our approach P, precision and recall are defined as $p = \frac{|T \cap P|}{|P|}$ and $r = \frac{|T \cap P|}{|T|}$, respectively. The results given in Table 2 are the average precision and recall values (over all bands in each of the ground truth sets M_c and M_f).

4.2 Upper Limits for Recall

Since the proposed approach relies on information that can be found on web pages, there exists an upper bound for the achievable performance. A band member that never occurs in the set of the 100 top-ranked web pages of a band obviously cannot be detected by our approach. As knowing these upper bounds is crucial to estimate the goodness of the recall values presented in Table 2,

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

Table 2. Overall precision and recall of the predicted (member, instrument)-pairs in percent for different query schemes and string distance functions on the ground truth sets M_c (upper table) and M_f (lower table). A filtering threshold of $t_{DF} = 0.25$ was used. The first value indicates the precision, the second the recall.

Precision/Recall on M_c			
	exact	*similar char*	*L2-JaroWinkler*
M	46.94 / 32.21	50.27 / 34.46	53.24 / 35.95
MR	42.49 / 31.36	45.42 / 33.86	48.20 / 35.32
MM	43.25 / 36.27	44.85 / 37.23	47.44 / 37.55
LUM	32.48 / 27.87	33.46 / 29.06	34.12 / 29.06
Precision/Recall on M_f			
	exact	*similar char*	*L2-JaroWinkler*
M	63.16 / 23.33	68.16 / 25.25	72.12 / 26.38
MR	52.42 / 21.33	55.63 / 23.12	59.34 / 24.82
MM	60.81 / 26.21	63.66 / 27.45	67.32 / 27.64
LUM	43.90 / 19.22	44.88 / 19.75	46.80 / 20.08

we analyzed how many of the actual band members given by the ground truth occur at least once in the retrieved web pages, i.e. for every band B, we calculate the recall, on the ground truth, of the N-grams extracted from B's web pages (without taking information about instruments into account). We verified that no band members were erroneously discarded in the N-gram selection phase. The results of these upper limit calculations using each query scheme and string matching function are depicted in Table 3 for both datasets M_c and M_f.

4.3 Influence of the Filtering Threshold t_{DF}

We also investigated the influence of the filtering threshold t_{DF} on precision and recall. Therefore, we conducted a series of experiments, in which we successively increased the value of t_{DF} between 0.0 and 1.0 with an increment of 0.01. The resulting precision/recall-plots can be found in Figures 1 and 2 for the ground truth datasets M_c and M_f, respectively. In these plots, only the results for exact string matching are presented for reasons of lucidity. Employing the other two, more tolerant, string distance functions just shifts the respective plots upwards. Since using low values for t_{DF} does not filter out many potential band members, the recall values tend to be high, but at the cost of lower precision. In contrast, high values of t_{DF} heavily prune the set of (member, instrument)-predictions and therefore generally yield lower recall and higher precision values.

4.4 Discussion of the Results

Taking a closer look at the overall precision and recall values given in Table 2 reveals that, for both datasets M_c and M_f, the query scheme M yields the highest precision values (up to more than 72% on the dataset M_f using Jaro-Winkler string matching), whereas the more specific scheme MM is able to

Table 3. Upper limits for the recall achievable on the ground truth datasets M_c (upper table) and M_f (lower table) using the 100 top-ranked web pages returned by *Google*. These limits are denoted for each of the search query scheme and string distance function. The values are given in percent.

Upper Limits for Recall on M_c			
	exact	*similar char*	*L2-JaroWinkler*
M	56.00	57.64	63.44
MR	50.28	53.53	60.92
MM	58.12	59.69	66.33
LUM	55.80	58.62	66.26
Upper Limits for Recall on M_f			
	exact	*similar char*	*L2-JaroWinkler*
M	52.97	55.15	62.01
MR	47.41	49.59	56.29
MM	56.40	57.62	64.08
LUM	55.21	57.27	64.11

achieve a higher recall on the ground truth (a maximum recall of nearly 38% on the dataset M_f using Jaro-Winkler string matching). The *LUM* scheme performs worst, independent of the used dataset and string distance function. The *MR* scheme performs better than *LUM*, but worse than *M* and *MM* with respect to both precision and recall.

Comparing the precision and recall values obtained using the dataset M_c with those obtained using M_f not surprisingly shows that for M_f the recall drops as this dataset contains more than double the number of band members as M_c and also lists members who spent a very short time with a band. For the same reasons, the precision is higher for the dataset M_f since obviously the chance to correctly predict a member is larger for a larger ground truth set of members.

Interestingly, comparing the upper limits for the recall for the two ground truth datasets (cf. Table 3) reveals that extending the set of the current band members with those who already left the band does not strongly influence the achievable recall (despite the fact that the number of band members in the ground truth set increases from 240 to 499 when adding the former members). This is a strong indication that the 100 top-ranked web pages of every band, which we use in the retrieval process, contain information about the current as well as the former band members to almost the same extent. We therefore conclude that using more than 100 web pages is unlikely to increase the quality of the (member, instrument)-predictions.

Regarding Figures 1 and 2, which depict the influence of the filtering parameter t_{DF} on the precision and recall values using the datasets M_c and M_f respectively, reveals that, for the dataset M_c, the query schemes *M*, *MR*, and *MM* do not stongly differ with respect to the achievable performance. Using the dataset M_f, in constrast, the results for the scheme *MR* are considerably worse than that for *M* and *MM*. It seems that album reviews (which are captured by the *MR* scheme) are more likely to mention the current band members than the

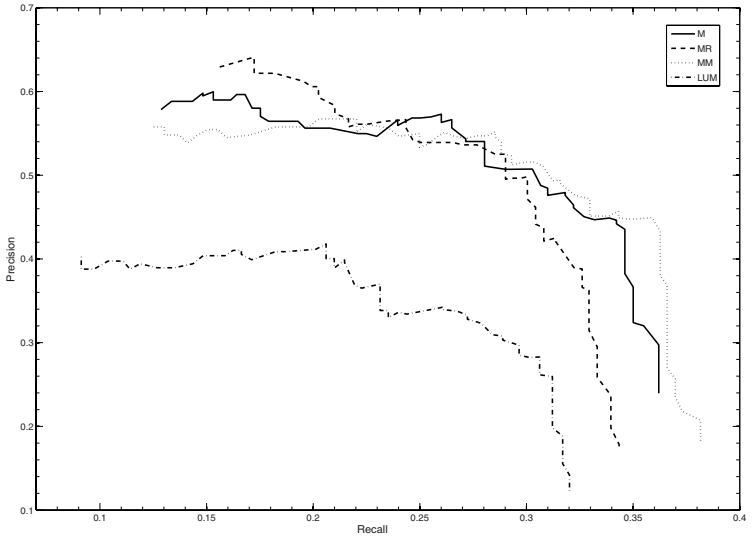

Fig. 1. Precision/recall-plot for the dataset M_c using the different query schemes and exact string matching

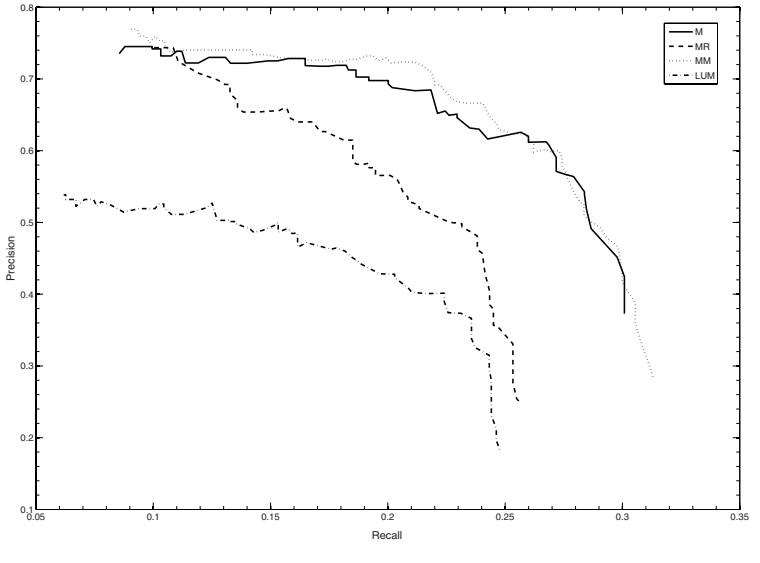

Fig. 2. Precision/recall-plot for the dataset M_f using the different query schemes and exact string matching

former ones. This explanation is also supported by the fact that the highest precision values on the dataset M_c are achieved with the *MR* scheme. Furthermore,

the precision/recall-plots illustrate the worse performance of the LUM scheme, independently of the filtering threshold t_{DF}.

To summarize, taking the upper limits for the recall into account (cf. Table 3), the recall values achieved with the proposed approach as given in Table 2 are quite promising, especially when considering the relative simplicity of the approach. Basically, the query scheme M yields the highest precision while the scheme MM yields the highest recall.

5 Conclusions and Future Work

We presented an approach to detecting band members and instruments they play within the band. To this end, we employ the techniques *N-gram extraction*, *named entity detection*, *rule-based linguistic analysis*, and *filtering according to document frequencies* on the textual content of the top-ranked web pages returned by *Google* for the name of the band under consideration. The proposed approach eventually predicts (member, instrument)-pairs. We evaluated the approach on two sets of band members from 51 bands, one containing the current members at the time this research was carried out, the other additionally including all former members. We presented and discussed precision and recall achieved for different search query schemes and string matching methods.

As for future work, we will investigate more sophisticated approaches to named entity detection. Employing machine learning techniques, e.g. to estimate the reliability of the rules used in the linguistic text analysis step, could also improve the quality of the results. We further aim at deriving complete band histories (by searching for dates when a particular artist joined or left a band), which would allow for creating time-dependent relationship networks. Under the assumption that bands which share or shared some members are similar to some extent, these networks could be used to derive a similarity measure. An application for this research is the creation of a domain-specific search engine for music artists, which is our ultimate aim.

Acknowledgments

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). The Austrian Research Institute for Artificial Intelligence acknowledges financial support by the Austrian ministries BMBWK and BMVIT.

References

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

2. Callan, J., Mitamura, T.: Knowledge-Based Extraction of Named Entities. In: Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM 2002), McLean, VA, USA, pp. 532–537. ACM Press, New York (2002)
3. Cano, P., Koppenberger, M.: The Emergence of Complex Network Patterns in Music Artist Networks. In: Proceedings of the 5th International Symposium on Music Information Retrieval (ISMIR 2004), Barcelona, Spain (October 2004)
4. Cimiano, P., Handschuh, S., Staab, S.: Towards the Self-Annotating Web. In: Proceedings of the 13th International Conference on World Wide Web (WWW 2004), pp. 462–471. ACM Press, New York, NY, USA (2004)
5. Cohen, W.W., Ravikumar, P., Fienberg, S.E.: A Comparison of String Distance Metrics for Name-Matching Tasks. In: Proceedings of the IJCAI 2003 Workshop on Information Integration on the Web (IIWeb 2003), Acapulco, Mexico, August 2003, pp. 73–78 (2003)
6. Goto, M., Goto, T.: Musicream: New Music Playback Interface for Streaming, Sticking, Sorting, and Recalling Musical Pieces. In: Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005), London, UK (September 2005)
7. Hearst, M.A.: Automatic Acquisition of Hyponyms from Large Text Corpora. In: Proceedings of the 14th Conference on Computational Linguistics, Nantes, France, August 1992, vol. 2, pp. 539–545 (1992)
8. Knees, P., Pohle, T., Schedl, M., Widmer, G.: A Music Search Engine Built upon Audio-based and Web-based Similarity Measures. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007), Amsterdam, The Netherlands, July 23-27 (2007)
9. Knees, P., Schedl, M., Pohle, T., Widmer, G.: An Innovative Three-Dimensional User Interface for Exploring Music Collections Enriched with Meta-Information from the Web. In: Proceedings of the ACM Multimedia 2006 (MM 2006), Santa Barbara, California, USA, October 23-26 (2006)
10. Kumaran, G., Allan, J.: Text Classification and Named Entities for New Event Detection. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004), pp. 297–304. ACM Press, New York, NY, USA (2004)
11. Mörchen, F., Ultsch, A., Nöcker, M., Stamm, C.: Databionic Visualization of Music Collections According to Perceptual Distance. In: Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005), London, UK (September 2005)
12. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. In: Proceedings of the Annual Meeting of the American Society for Information Science (ASIS 1998), January 1998, pp. 161–172 (1998)
13. Schedl, M., Knees, P., Widmer, G.: Discovering and Visualizing Prototypical Artists by Web-based Co-Occurrence Analysis. In: Proceedings of the Sixth International Conference on Music Information Retrieval (ISMIR 2005), London, UK (September 2005)
14. Schedl, M., Pohle, T., Knees, P., Widmer, G.: Assigning and Visualizing Music Genres by Web-based Co-Occurrence Analysis. In: Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR 2006), Victoria, Canada (October 2006)

15. Shinyama, Y., Sekine, S.: Named Entity Discovery Using Comparable News Articles. In: Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004), Morristown, NJ, USA. Association for Computational Linguistics, p. 848 (2004)
16. Vignoli, F., van Gulik, R., van de Wetering, H.: Mapping Music in the Palm of Your Hand, Explore and Discover Your Collection. In: Proceedings of the 5th International Symposium on Music Information Retrieval (ISMIR 2004), Barcelona, Spain (October 2004)
17. Zadel, M., Fujinaga, I.: Web Services for Music Information Retrieval. In: Proceedings of the 5th International Symposium on Music Information Retrieval (ISMIR 2004), Barcelona, Spain (October 2004)