Towards an Automatically Generated Music Information System via Web Content Mining

Markus Schedl[1], Peter Knees[1], Tim Pohle[1], and Gerhard Widmer[1,2]
markus.schedl@jku.at

[1] Department of Computational Perception
Johannes Kepler University
Linz, Austria
http://www.cp.jku.at

[2] Austrian Research Institute for Artificial Intelligence
Vienna, Austria
http://www.ofai.at

Abstract. This paper presents first steps towards building a music information system like *last.fm*, but with the major difference that the data is automatically retrieved from the WWW using web content mining techniques. We first review approaches to some major problems of music information retrieval (MIR), which are required to achieve the ultimate aim, and we illustrate how these approaches can be put together to create the *automatically generated music information system* (AGMIS). The problems addressed in this paper are *similar and prototypical artist detection*, *album cover retrieval*, *band member and instrumentation detection*, *automatic tagging of artists*, and *browsing/exploring web pages related to a music artist*. Finally, we elaborate on the currently ongoing work of evaluating the methods on a large dataset of more than 600,000 music artists and on a first prototypical implementation of AGMIS.

1 Introduction and Context

Music information systems like *last.fm* [1] typically offer multimodal information about music artists, albums, and tracks (e.g. genre and style, similar artists, biographies, song samples, or images of album covers). In common music information systems, such information is usually gained and revised by experts (e.g. *All Music Guide* [2]), or relies on user participation (e.g. *last.fm*). In contrast, we are building such a system by automatically extracting the required information from the web.

Automatically retrieving descriptive information about music artists is an important task in music information retrieval (MIR) as it allows for enriching music players [13], for automatic biography generation [4], for enhancing user interfaces to browse music collections [7, 6], or for defining similarity measures between artists, a key concept in MIR. Similarity measures enable, for example, creating relationship networks [10], building music recommender systems [15] or music search engines [5].

In the following, we first give a brief overview of the available techniques which we are refining at the moment. Hereafter, we present the currently ongoing work of combining these techniques to build the *automatically generated music information system* (AGMIS).

2 Mining the Web for Music Artist-Related Information

All of the applied methods rely on the availability of artist-related data in the WWW. Our principal approach to extracting such data is the following. Given only a list of artist names, we first query a search engine[3] to retrieve the URLs of up to 100 top-ranked search results for each artist. The content available at these URLs is extracted and stored for further processing. To overcome the problem of artist or band names that equal common speech words and to direct the search towards the desired information, we use task-specific query schemes, like *"band name"+music+members* to obtain data related to band members and instrumentation. Depending on the task to solve, we then create either a document-level inverted file index or a word-level index [16]. In some cases, we use a special dictionary of musically relevant terms to perform indexing. After having indexed the web pages, we gain artist-related information of various kinds as described in the following.

2.1 Relations between Artists

A key concept in music information retrieval and crucial part of any music information system are *similarity relations* between artists. To model such relations, we use an approach that is based on co-occurrence analysis [9]. More precisely, the similarity between two artists a and b is defined as the conditional probability that the artist name a occurs on a web page that was returned as response to the search query for the artist name b and vice versa, formally $\frac{1}{2} \cdot \left(\frac{df_{a,B}}{|B|} + \frac{df_{b,A}}{|A|} \right)$, where A represents the set of web pages returned for artist a and $df_{a,B}$ is the document frequency of the artist name a calculated on the set of web pages returned for artist b. Having calculated the similarity for each pair of artists in the artist list, we can output, for any artist, a list of most similar ones. Evaluation in an artist-to-genre classification task on a set of 224 artists from 14 genres yielded accuracy values of about 85%.

Co-occurrences of artist names on web pages (together with genre information) can also be used to derive information about the *prototypicality of an artist for a certain genre* [10, 11]. To this end, we make use of the asymmetry of the co-occurrence-based similarity measure.[4] We developed an approach that is based on the forward link/backlink-ratio of two artists a and b from the same genre, where a backlink of a from b is defined as any occurrence of artist a on a web page that is known to contain artist b, whereas a forward link of a to b is defined as any occurrence of b on a web page known to mention a. Relating the number of forward links to the number of backlinks for each pair of artists from the same genre, a ranking of the artist prototypicality for the genre under consideration is obtained. A more extensive description of the approach can be found in [11].

2.2 Band Member and Instrumentation Detection

Another type of information indispensible for a music information system is *band members and instrumentation*. In order to capture such aspects, we first apply a named entity detection approach that basically relies on extracting N-grams and on filtering w.r.t. capitalization and words contained in the *iSpell English Word Lists* [3]. The remaining N-grams are regarded as potential band members.

[3] We commonly used *Google* in our experiments, but also experimented with *exalead*.

[4] In general, $\frac{df_{a,B}}{|B|} \neq \frac{df_{b,A}}{|A|}$.

Subsequently, we perform linguistic analysis to obtain the actual instrument(s) of each member. To this end, a set of seven patterns like *"M plays the I"*, where M is the potential member and I is the instrument, is applied to the N-grams (and the surrounding text as necessary). The document frequencies of the patterns are recorded and summed up over all seven patterns for each (M, I)-tuple. After having filtered out those (M, I)-pairs whose document frequency is below a dynamically adapted threshold in order to suppress uncertain information, the remaining (M, I)-tuples are predicted for the band under consideration. More details as well as an extensive evaluation can be found in [14].

2.3 Automatic Tagging of Artists

For automatically attributing textual descriptors to artists, we use a dictionary of about $1,500$ musically relevant terms to index the web pages. As for term weighting, three different measures (document frequency, term frequency, and $TF \times IDF$) were evaluated in a yet unpublished quantitative user study. This study showed, quite surprisingly, that the simple document frequency measure outperformed the well-established $TF \times IDF$ measure significantly (according to Friedman's non-parametric two-way analysis of variance). Thus, for the AGMIS, we will probably use this measure to automatically select the most appropriate tags for each artist.

2.4 Co-Occurrence Browser

To easily access the top-ranked web pages of any artist, we designed a user interface, which we call the *Co-Occurrence Browser* (COB). Based on the dictionary used for automatic tagging, the COB groups the web pages of the artist under consideration w.r.t. the document frequencies of co-occurring terms. These groups are then visualized using the approach presented in [12]. Thus, the COB allows for browsing the artist's web pages by means of descriptive terms. Furthermore, the multimedia content present on the web pages is extracted and made available via the user interface.

2.5 Album Cover Retrieval

Preliminary attempts to automatically retrieve album cover artwork were made in [8]. We refined the methods presented in this paper and conducted experiments with content-based as well as context-based methods for detecting images of album covers. We found that using the text distance between album names and *img*-tags in the HTML file at character level gives a quite good indication whether an image is an album cover or not. The results could further be improved by rejecting images that have non-quadratic dimensions or appear to show a scanned disc (which happens quite often). On a challenging collection of about $3,000$ albums, we estimated a precision of approximately 60%.

3 Building the AGMIS

Currently, our work is focusing on the large-scale retrieval of artist-related information and on building a prototypical implementation of the AGMIS user interface. As for retrieval, the search engine *exalead* was used to obtain a list of more than $26,000,000$ URLs (for a total of $600,000$ artists from 18 genres). We are fetching these URLs using a self-made, thread-based Java program that offers load balancing between the destination hosts. A file index of the retrieved

web documents will be build subsequently.
As for the user interface, Figure 1 shows a sample web page created by a prototypical implementation of AGMIS (based on Java Servlet and Applet technologies). This prototype incorporates the information whose extraction and presentation was described in Section 2. On the left-hand side, textual information about the artist *Hammerfall* is offered to the user, whereas on the right, the user interface of the COB is embedded as a Java Applet. The page is further enriched by displaying images of album covers in its lower part (which are omitted in the screenshot due to copyright reasons).

4 Conclusions and Future Work

We presented a set of methods that address current problems in the field of web-based music information retrieval and showed how we will apply them to create an automatically generated music information system, which we call AGMIS.
Future work will mainly focus on evaluating the presented approaches on the large corpus which we are currently building. After having fetched them, we will look into efficient methods for high-speed indexing of the retrieved web pages and for organizing and storing the information extracted from the index via the approaches presented in Section 2. Finally, the user interface for accessing the music information will probably need some updates.

Acknowledgments

This research is supported by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung (FWF) under project number L112-N04.

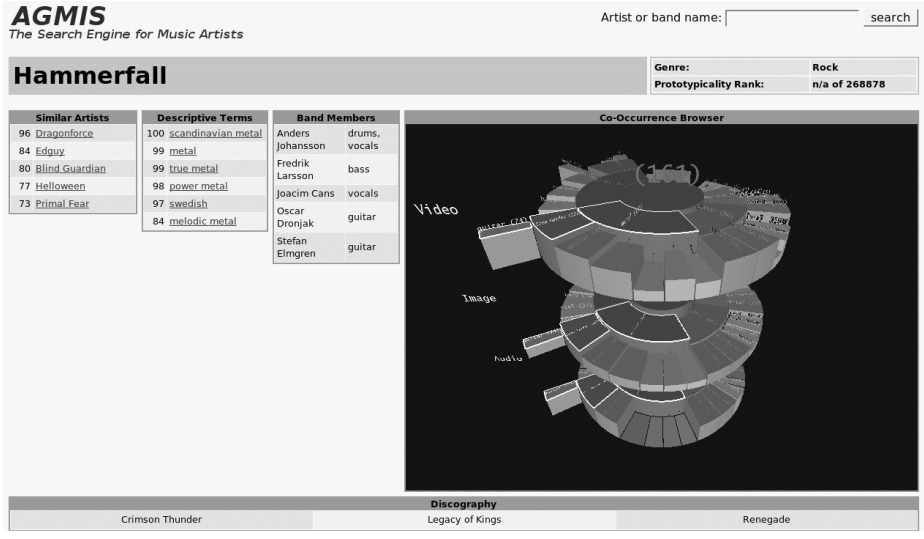

Fig. 1. Screenshot of a prototypical implementation of AGMIS.

References

1. http://last.fm, 2007. last access: 2007-10-04.
2. http://www.allmusic.com, 2007. last access: 2007-10-04.
3. http://wordlist.sourceforge.net, 2007. last access: 2007-10-04.
4. Harith Alani, Sanghee Kim, David E. Millard, Mark J. Weal, Wendy Hall, Paul H. Lewis, and Nigel R. Shadbolt. Automatic Ontology-Based Knowledge Extraction from Web Documents. *IEEE Intelligent Systems*, 18(1):14–21, 2003.
5. Peter Knees, Tim Pohle, Markus Schedl, and Gerhard Widmer. 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 2007.
6. Peter Knees, Markus Schedl, Tim Pohle, and Gerhard Widmer. An Innovative Three-Dimensional User Interface for Exploring Music Collections Enriched with Meta-Information from the Web. In *Proceedings of the 14th ACM Conference on Multimedia 2006*, Santa Barbara, CA, USA, October 2006.
7. Elias Pampalk and Masataka Goto. MusicSun: A New Approach to Artist Recommendation. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, Vienna, Austria, September 2007.
8. Markus Schedl, Peter Knees, Tim Pohle, and Gerhard Widmer. Towards Automatic Retrieval of Album Covers. In *Proceedings of the 28th European Conference on Information Retrieval (ECIR 2006)*, London, UK, April 2006.
9. Markus Schedl, Peter Knees, and Gerhard Widmer. A Web-Based Approach to Assessing Artist Similarity using Co-Occurrences. In *Proceedings of the 4th International Workshop on Content-Based Multimedia Indexing (CBMI 2005)*, Riga, Latvia, June 2005.
10. Markus Schedl, Peter Knees, and Gerhard Widmer. Discovering and Visualizing Prototypical Artists by Web-based Co-Occurrence Analysis. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, London, UK, September 2005.
11. Markus Schedl, Peter Knees, and Gerhard Widmer. Investigating Web-Based Approaches to Revealing Prototypical Music Artists in Genre Taxonomies. In *Proceedings of the 1st IEEE International Conference on Digital Information Management (ICDIM 2006)*, Bangalore, India, December 2006.
12. Markus Schedl, Peter Knees, Gerhard Widmer, Klaus Seyerlehner, and Tim Pohle. Browsing the Web Using Stacked Three-Dimensional Sunbursts to Visualize Term Co-Occurrences and Multimedia Content. In *Proceedings of the 18th IEEE Visualization 2007 Conference (Vis 2007)*, Sacramento, CA, USA, October 2007.
13. Markus Schedl, Tim Pohle, Peter Knees, and Gerhard Widmer. 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.
14. Markus Schedl and Gerhard Widmer. Automatically Detecting Members and Instrumentation of Music Bands via Web Content Mining. In *Proceedings of the 5th Workshop on Adaptive Multimedia Retrieval (AMR 2007)*, Paris, France, July 2007.
15. Mark Zadel and Ichiro Fujinaga. Web Services for Music Information Retrieval. In *Proceedings of the 5th International Symposium on Music Information Retrieval (ISMIR 2004)*, Barcelona, Spain, October 2004.
16. Justin Zobel and Alistair Moffat. Inverted Files for Text Search Engines. *ACM Computing Surveys*, 38(2):6, 2006.