Automatically Describing Music on a Map

Peter Knees[1], Tim Pohle[1], Markus Schedl[1], and Gerhard Widmer[1,2]

[1] Dept. of Computational Perception, Johannes Kepler University Linz, Austria
[2] Austrian Research Institute for Artificial Intelligence (OFAI)
peter.knees@jku.at

Abstract. In this paper, we present a technique to automatically create music maps labeled with semantic descriptors, the so called *Music Description Maps (MDM)*. Based on a Self-organizing Map (SOM) trained on audio features, we create term profiles that characterize the type of music in the various clusters. To this end, we efficiently retrieve music-related term descriptors for music artists from the Web. These descriptors are used in conjuction with a SOM-labeling strategy to identify words and phrases commonly used in the context of the associated music. Additionally, regions of similar clusters are uncovered. Music maps labeled in such a manner can aid the user in retrieving desired music from a very large repository, either by providing landmarks on the map or by allowing the formulation of queries consisting of terms describing the musical content.

1 Introduction

Since digital music collections comprise vast amounts of pieces nowadays, automatic structuring and organization of large music repositories is a central challenge. To allow automatic organization, features to describe music have to be available. The most common approach to acquire these features is to extract them directly from the audio signal (for an overview, see e.g. [1, 2]). While this approach is capable of modelling certain sound characteristics, it can not capture other information relevant to the perception of music, e.g. contextual or social factors. A complementary approach to derive features about music is to exploit meta-data, e.g. by analyzing texts about music or musical artists and their work extracted from the Web. Such meta-data can be used to calculate, for example, artist similarity by applying text retrieval methods, e.g. [3–5]. However, while this approach is capable of capturing social factors to a certain extent and performs well in tasks at the artist level, it has not yet been applied successfully to calculate similarities at the track level, mainly because of the variable number of related Web pages for different tracks.

With this paper, we aim at bridging the semantic gap between signal-based organization of music archives at the level of individual tracks and semantic descriptions of the work of musical artists. We utilize a Self-organizing Map (SOM) [6] trained on audio features and label it with terms that describe the musical characteristics on the different map units. The music labels are obtained

by querying Google with the names of the corresponding artists and applying a commonly used SOM labeling strategy. Furthermore, we propose a technique to uncover and fuse coherent (or at least similar) regions on the maps in order to make them more clear for the user. The resulting *Music Description Map (MDM)* can be very useful for browsing large music repositories, since the terms serve as landmarks on the map, allowing better orientation. Furthermore, the MDM provides a mapping between musical pieces in the audio feature space and the cultural context. This mapping could also be utilized to query a music retrieval system by describing musical contents with familiar terms.

2 Related Work

Related work comprises basically three topics: Clustering approaches to structuring music repositories, labeling strategies for text collections, and approaches to combining audio-based features with information derived from the Web.

Most systems that create a map for music organization purposes use a SOM to form clusters of similar songs and to project them on a 2-dimensional plane with the goal of providing intuitive access to a collection. The first approach that incorporated SOMs to structure music collections is presented in [7]. This approach has been modified by Pampalk to create the *Islands of Music* interface [8, 9]. In addition to this approach, several extensions have been proposed, e.g. the usage of Aligned SOMs [10] to enable a seamless shift of focus between different aspects of similarity or a hierarchical component to cope with very large music collections [11]. In [12], SOMs are utilized for browsing in collections and intuitive playlist generation on portable devices. In [13], we extended the *Islands of Music* approach to provide a three-dimensional game-like virtual reality landscape with sound auralization in which the user can freely navigate using a gamepad. Other approaches use SOM derivatives [14] or similar techniques like FastMap [15]. In [16], textual information from *Amazon* reviews is used to structure music collections by incorporating a fixed list of musically related terms to describe similar artists. In [17], hierachical one-dimensional SOMs are used to guide the user to relevant artists. At each level, the user chooses from sets of music descriptions that are determined via term selection approaches (see below). In this paper, we aim at overcoming the limitation to the artist level by combining audio-based clustering at the track level with artist-related features from the Web.

To determine useful descriptors for clusters in text collections, several strategies have been proposed. One approach is the SOM-labeling technique by Lagus and Kaski which was developed to support the WEBSOM technique [18, 19]. Another approach to find descriptive terms for text documents clustered using a SOM is the LabelSOM approach [20]. While the Lagus and Kaski method determines the importance of descriptors for each cluster based on the contained items, LabelSOM chooses those terms that represent the most relevant dimensions for assigning data to a cluster in the training phase. As a consequence, the Lagus and Kaski approach can also be used in "situations where the data of interest is numeric, but where some texts can be meaningfully associated with the

data items", as the authors state in the conclusions of [18]. Hence, we can also apply this strategy to label a SOM trained on audio features with semantic descriptors extracted automatically from the Web, as we will demonstrate in Section 4.

Finally, we review approaches that incorporate both signal-based and metadata-based methods. In [21], audio-based and web-based genre classification are used for the task of style detection. [22] linearly combines audio-based track similarity with Web-based artist similarity to obtain a new similarity measure. In [23], we exploit Web-based artist similarity to reduce the number of necessary distance calculations between audio tracks for automatic playlist generation. In [24], Whitman uses audio features and semantic descriptors to learn the meaning of certain acoustic properties and to overcome the semantic gap.

In contrast, in this paper, we pursue a top-down approach to find a mapping between signal and meaning. Instead of assigning descriptions to certain low-level characteristics of an audio signal, we aim at describing consistent groups of musical pieces, i.e. regions on a map, with culturally related terms.

3 Creating Music Maps

Although the Music Description Map (see Section 4) is not bound to a specific audio similarity measure, in this section we briefly describe the steps we perform to obtain a music map.

The first step is to calculate pairwise similarities between all tracks based on their audio signal. As pointed out in [25, 2], it is possible to accelerate the algorithm described in [26] by a factor of about 20, while the classification accuracy remains almost the same. One track is described by the mean and the full covariance matrix computed from *Mel Frequency Cepstral Coefficients* (MFCCs) over all audio frames of the song. The models of two songs are compared by calculating a modified Kullback-Leibler distance on the means and covariance matrices [25]. As the resulting distance matrix only contains the pairwise distances between tracks, and the SOM algorithm expects points in Euclidean space as input, we interpret each column of the similarity matrix as one vector in Euclidean space, where the i^{th} row corresponds to the i^{th} song. However, the feature extraction process described produces a distance matrix that is not well scaled for this purpose. Although it is possible to use a normalization as proposed in [25, 2], we opted for a different approach that has the advantage of being parameterless. Applying a method called *Proximity Verification* [27], we replace the distances in the distance matrix D by a rank-based measure. The entries of each row of the distance matrix D are sorted in ascending order, and each original entry of the row is replaced with its rank. The resulting distance matrix (denoted D_1 here) is transformed into the final matrix by adding the transpose (resulting in a symmetric matrix): $D_{final} := D_1 + D_1'$. The resulting matrix has a better distribution of distances than the original distance matrix, and seems to be better suited as input for the SOM algorithm. To create the actual map, we train a SOM using the Linear initialization method. An example of a resulting music map is depicted in Fig. 1.

4 The Music Description Map

To create an MDM, which describes the kind of music in the different regions of the map, we have to perform three steps. First, we have to retrieve information from the Web to create term profile descriptions of the musical artists contained in the collection. In a second step, we associate each track in the collection with the term characterization of its corresponding artist and label the SOM based on these representations. Finally, we search for similarly labeled clusters to detect larger coherent regions on the MDM.

4.1 Artist Term Profile Retrieval

While it is difficult to find specific information on certain songs, extracting information describing the general style of an artist is feasible. Usually [3, 4, 17], the acquisation of artist descriptors is realized by invoking Google with a query like `‘‘artist name’’ music review` and analyzing the first 50 returned pages, e.g. by counting term frequency (tf) and document frequency (df) for either single words, bigrams, or trigrams and combine them into the well known $tf \times idf$ measure. However, downloading 50 pages for each artist is bandwidth and time consuming. To speed up the artist profile extraction, which is crucial to allow for integration of the technique also in time-critical applications like [13], we simplify the search for musical style by formulating the query `‘‘artist name’’ music style` and retrieving Google's result page containing links to the first 100 pages and extracts of the relevant sections ("snippets"). Instead of downloading each of the returned sites, we directly analyze the complete result page, i.e. the snippets presented. Thus, we can reduce the effort to downloading and analyzing only one web page per artist. Another advantage of analyzing only the "digest" of the artist-related pages is to incorporate only information from the most important sections of the Web pages, i.e. the most relevant sections with respect to the query. Otherwise, structural analysis of each Web page would be necessary to avoid inclusion of unrelated text portions (e.g. from navigation bars). To eliminate totally unrelated words, we use a reduced version of the dictionary used in [17]. Thus, we only count occurrences of words or phrases that are contained in this dictionary of music-related terms. After obtaining a term frequency representation of the dictionary vector for each artist, we determine the important terms for each cluster as described next.

4.2 SOM Labeling

Once we have gathered music term vectors for all artists, we are in need of a strategy to determine those words that discriminate between the music in one region of the map and music in another (e.g. *Music* is not a discriminating word, since it occurs very frequently for all artists; *Piano* would be a valuable word to indicate piano music, assuming piano music forms a distinct cluster on the map).

We decided to apply the SOM-labeling strategy proposed by Lagus and Kaski [18] (cf. Section 2). In their heuristically motivated weighting scheme,

Dance (7) Hip-Hop (3) Jazz Metal Punk	Dance (13) Hip-Hop (2) Jazz Pop	Dance (23) Hip-Hop (2) Jazz	Dance (15) Hip-Hop (4) Pop (4) Jazz (3)	Dance (144) Pop (113) Hip-Hop (50) Jazz (19) Punk (15) Metal (10)	Dance (14) Pop (6) Punk (3) Hip-Hop (2) Jazz (2) Metal (2)	Dance (61) Pop (23) Metal (13) Punk (8) Hip-Hop
Dance (4) Hip-Hop Pop	Dance	Dance (6) Hip-Hop Pop	Dance (2)	Dance (3) Metal Pop	Pop (3) Metal Punk	Dance (28) Pop (17) Metal (11) Punk (7) Hip-Hop (2) Jazz
Dance (7) Metal	Dance (2) Punk	Dance (5) Hip-Hop (2) Pop	Dance (3) Metal Pop Punk	Dance (6) Hip-Hop Pop	Metal Punk	Metal (23) Pop (21) Punk (14) Dance (12) Hip-Hop (2)
Dance (6) Hip-Hop	Dance (3)	Hip-Hop (2) Dance Metal Punk	Dance (2) Punk	Dance (4)	Punk (3) Metal	Metal (42) Punk (18) Pop (12) Dance (10) Hip-Hop (2)
Jazz		Punk	Metal (94) Punk (37) Dance (3) Pop (3)	Metal (2) Dance	Metal (3) Punk (3)	Metal (192) Punk (173) Pop (6) Dance (4) Hip-Hop
Jazz		Punk	Punk			Metal (2) Pop (2) Punk (2)
Metal (3) Jazz	Jazz (2)					Punk (4) Metal (2) Pop (2)
Classical (217) Jazz (129) Metal (3)	Jazz (3) Metal	Jazz (5) Hip-Hop	Dance (2) Pop (2)	Metal	Metal (2)	Punk (5) Pop (4) Metal
Jazz (23) Classical (4) Pop (2) Dance	Jazz (2) Dance	Dance (2)	Hip-Hop (3) Jazz (2) Pop (2) Dance	Hip-Hop (2) Pop		Pop (4) Punk (2) Hip-Hop Jazz
Jazz (123) Dance (10) Pop (6) Classical (5) Hip-Hop (3) Metal (3)	Jazz (30) Dance (7) Pop (4) Hip-Hop (3)	Jazz (24) Pop (14) Dance (10) Hip-Hop (6) Metal	Pop (59) Jazz (55) Hip-Hop (37) Dance (20) Metal (2) Punk	Hip-Hop (51) Pop (44) Dance (30) Jazz (21) Metal	Hip-Hop (34) Pop (26) Dance (25) Jazz (5) Metal (2)	Hip-Hop (23) Dance (12) Pop (12) Jazz (4) Metal (3) Punk

Fig. 1. A 7×10 SOM trained on a collection containing 2 572 tracks (by 331 artists) assigned to 7 genres: Classical, Dance, Hip-Hop, Jazz, Metal, Pop, and Punk. Tracks from Jazz, as well as from Classical, tend to cluster at the (lower) left, Dance at the top. Center and right side are dominated by Punk and Metal. Hip-Hop is mainly found on the bottom. Pop occurs frequently in conjunction with Dance and Hip-Hop.

knowledge of the strucure of the SOM is exploited to enforce the emergence of areas with coherent descriptions. To this end, terms from directly neighboring units are accumulated and terms from a more distant "neutral zone" are ignored. The goodness G^2 of a term t as a descriptor for unit u is calculated as

$$G^2(t,u) = \frac{\left[\sum_{k \in A_0^u} F(t,k)\right]^2}{\sum_{i \notin A_1^u} F(t,i)}, \tag{1}$$

where $k \in A_0^u$ if the (Manhattan) distance of units u and k on the map is below a threshold r_0, and $i \in A_1^u$ if the distance of u and i is greater than r_0 and smaller than r_1 (in our experiments we set $r_0 = 1$ and $r_1 = 2$). $F(t,u)$ denotes the relative frequency of term t on unit u and is calculated as

$$F(t,u) = \frac{\sum_a f(a,u) \cdot tf(t,a)}{\sum_v \sum_a f(a,u) \cdot tf(v,a)}, \tag{2}$$

where $f(a,u)$ gives the number of tracks of artist a on unit u and $tf(t,a)$ the term frequency of term t for artist a. We ignore all entries with $G^2 < 0.01$ and select at most 30 terms to appear on a map unit (provided that there is enough space to display them). Furthermore, we set the font size of a term according to its score. However, this approach leads to very cluttered maps. Additionally, many neighboring units contain very similar descriptions. Thus, one could easily happen "not to see the wood for the trees" when orienting on the map. Since we aim at providing clearly arranged maps to make it *simpler* to find music, we try to find coherent parts of the MDM and join them to single clusters.

4.3 Connecting Similar Clusters

To find adjacent units that have similar descriptors we apply the following heuristic. First, we sort all units according to the maximum G^2 values of the contained terms. Starting with the highest ranked unit, we perform a recursive cluster expansion step for all units. In this step, we try to find similarly labeled units among the adjacent four neighbors. The idea is to create one vector representation that adequately reflects the vectors of the contained units. We achieve this by comparing the *Cosine normalized* description vectors of both units with the Cosine normalized vector obtained by adding both vectors. Both normalized unit vectors are compared with the normalized "sum vector" by calculating the *Euclidean distance*. If both distances are below a threshold d (we use an empirically determined value of 0.4), i.e. if the resulting vector is similar to the original ones and thus capable to sufficiently represent both original vectors, we admit the candidate unit to the cluster and assign the sum of the unnormalized vectors to all units in the cluster. Thus, the larger the regions grow, the more important become its descriptors. For all absorbed units, this procedure is repated recursively. An example of an MDM with connected map units can be found in Fig. 2.

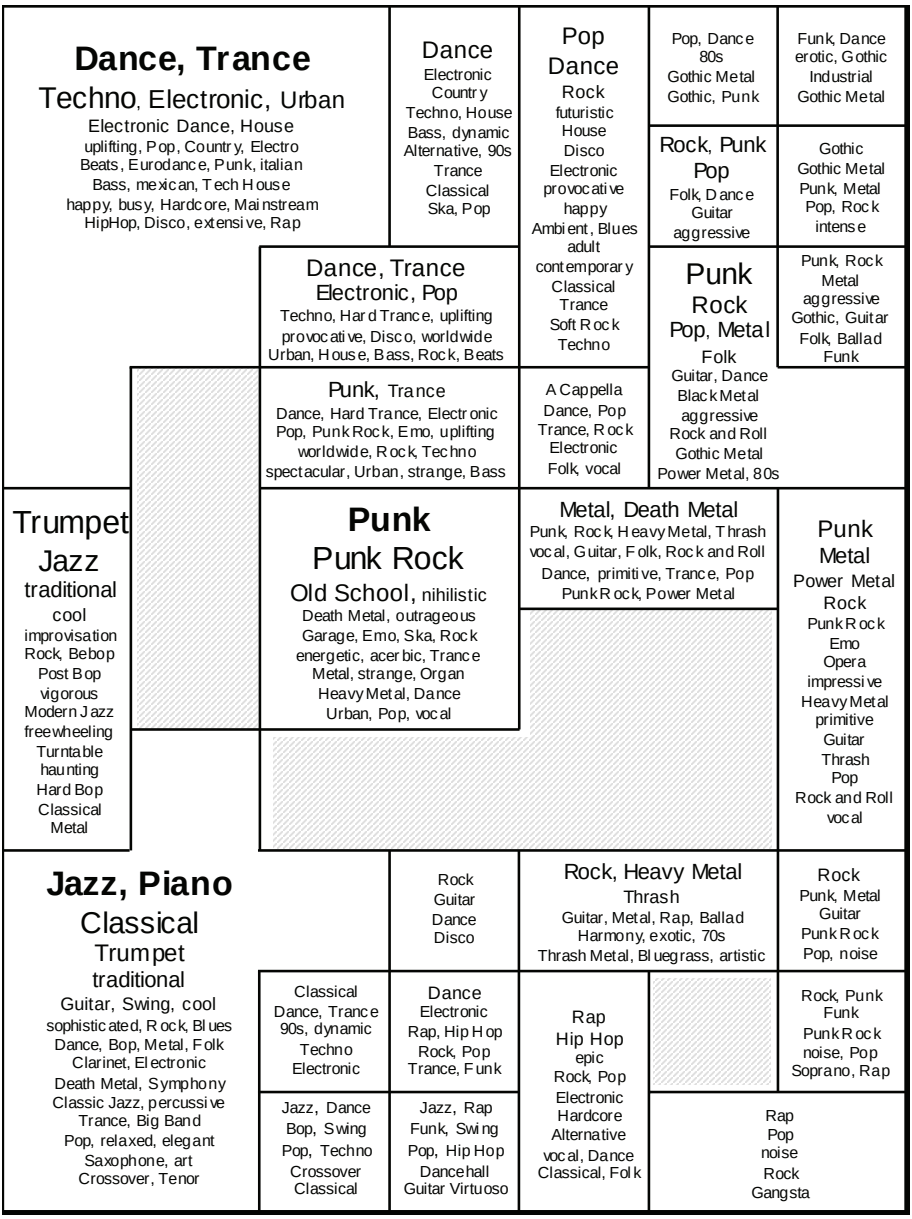

Fig. 2. A Music Description Map with joined similar clusters. The displayed terms describe the musical style in the different regions of the underlying music map (7×10 SOM) from Fig. 1. The size of the terms reflect the importance of the descriptor for the corresponding region. For reasons of readability, the linebreaks have been edited manually.

5 Evaluation

Taking a look at Fig. 2, one can easily identify the contained music at first sight or at least get a good impression of the contained collection (2 572 tracks from 7 genres). A more detailed examination reveals that Jazz has two major clusters – one consisting of *Piano* Jazz (together with Classical Music), and one consisting of *Trumpet* Jazz. Also genres like Metal are represented through more distinct information (e.g. *Gothic* vs. *Power Metal*). The contained adjectives (*energetic*, *percussive*, *aggressive*, etc.) can also give information to users unfamiliar with the presented style descriptors.

Evaluating such an approach quantitatively is rather difficult, since we do not have access to any form of ground truth, i.e. any pre-labeled corpus of music pieces. Hence, we decided to ask 6 users to provide us with (a small selection) of music pieces from their personal collection, i.e. music they are familiar with. Based on every collection, we created a small MDM (6×4), which we presented to the corresponding user together with a list of the contained music pieces. The users were then asked to assign each track to the cluster that best describes each track in their opinion. In case of uncertainty, it was also possible to select a second best matching cluster. The results can be found in Table 1.

test person	1	2	3	4	5	6	*total*
tracks in collection	54	35	28	45	51	41	254
clusters on MDM	8	5	13	8	6	12	
matching assignments (1st choice)	21	18	7	10	42	7	105
matching assignments (2nd choice)	5	4	n.a.	6	3	0	18
matching assignments (total)	26	22	7	16	45	7	123
percent	48.1	62.9	25.0	35.6	88.2	17.1	48.4

Table 1. Evaluation results of track to MDM cluster assignment.

Obviously, the results are very heterogeneous. At first glance, a high number of emerging clusters seems to be responsible for bad results. A deeper investigation reveals that both, high number of clusters and bad results, have the same sources, namely many non-english music pieces and many outliers. In test case 6, the collection basically consisted solely of Rap and Dance music with strong beats and all clusters on the map where labeled very similarly. In contrast, collections that contained consistent subsets of music (which could also be identified by the audio measure) led to few large, clearly separated clusters. On these collections, highest matchings between MDM and user opinions could be observed.

6 Conclusions and Future Work

We presented the Music Description Map (MDM), which is an approach to automatically create music maps labeled with semantic descriptors by applying a SOM-labeling strategy to Web-based artist term profiles. As can be seen in Fig. 2, in most cases, terms describing the style and genre of music are most important to describe the content of a cluster.

However, experiments with users showed that there is still ample space for improvements. While well known music is substantially represented on the Web and can also be sufficiently captured by the used dictionary, many non-english music styles can not be described and result in misleading terms. Furthermore, also outliers impose some problems on the labeling. In fact, the MDM assumes a "perfect" underlying similarity measure and clustering, meaning that all clusters contain similar music and no outliers (although it is clear that this will never be satisfied in practice).

For future work, we will try to improve the quality by modifying the SOM-labeling to better reflect the number of pieces in the clusters and using multi-language dictionaries to capture more types of music.

7 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. Pohle, T.: Extraction of Audio Descriptors and Their Evaluation in Music Classification Tasks. Master's thesis, Technical University of Kaiserslautern, DFKI, OFAI, http://kluedo.ub.uni-kl.de/volltexte/2005/1881/ (2005)
2. Pampalk, E.: Computational Models of Music Similarity and their Application to Music Information Retrieval. PhD thesis, Vienna University of Technology (2006)
3. Whitman, B., Lawrence, S.: Inferring Descriptions and Similarity for Music from Community Metadata. In: Proc. of the 2002 International Computer Music Conference, Goeteborg, Sweden (2002)
4. Knees, P., Pampalk, E., Widmer, G.: Artist Classification with Web-based Data. In: Proc. of 5th International Conference on Music Information Retrieval (ISMIR'04), Barcelona, Spain (2004)
5. Cohen, W., Fan, W.: Web-Collaborative Filtering: Recommending Music by Crawling The Web. WWW9 / Computer Networks **33**(1-6) (2000)
6. Kohonen, T.: Self-Organizing Maps. 3rd edn. Springer, Berlin (2001)
7. Rauber, A., Frühwirth, M.: Automatically Analyzing and Organizing Music Archives. In: Proc. of the 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL'01), Darmstadt, Germany (2001)
8. Pampalk, E.: Islands of Music: Analysis, Organization, and Visualization of Music Archives. Master's thesis, Vienna University of Technology (2001)
9. Pampalk, E., Rauber, A., Merkl, D.: Content-Based Organization and Visualization of Music Archives. In: Proc. of the ACM Multimedia, Juan les Pins, France, ACM (2002)
10. Pampalk, E., Dixon, S., Widmer, G.: Exploring Music Collections by Browsing Different Views. Computer Music Journal **28**(2) (2004)

11. Schedl, M.: An Explorative, Hierarchical User Interface to Structured Music Repositories. Master's thesis, Vienna University of Technology (2003)
12. Neumayer, R., Dittenbach, M., Rauber, A.: PlaySOM and PocketSOMPlayer, Alternative Interfaces to Large Music Collections. In: Proc. of the 6th International Conference on Music Information Retrieval (ISMIR'05), London, UK (2005)
13. 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: Proc. of the ACM Multimedia 2006, Santa Barbara, California, USA (2006)
14. Mörchen, F., Ultsch, A., Nöcker, M., Stamm, C.: Databionic visualization of music collections according to perceptual distance. In: Proc. of the 6th International Conference on Music Information Retrieval (ISMIR'05), London, UK (2005)
15. Cano, P., Kaltenbrunner, M., Gouyon, F., Batlle, E.: On the Use of Fastmap for Audio Retrieval and Browsing. In: Proc. of the International Conference on Music Information Retrieval (ISMIR'02), Paris, France (2002)
16. Vembu, S., Baumann, S.: A Self-Organizing Map Based Knowledge Discovery for Music Recommendation Systems. In: Proc. of the 2nd Internation Symposium on Computer Music Modeling and Retrieval (CMMR'04), Esbjerg, Denmark (2004)
17. Pampalk, E., Flexer, A., Widmer, G.: Hierarchical Organization and Description of Music Collections at the Artist Level. In: Proc. of the 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL'05), Vienna, Austria (2005)
18. Lagus, K., Kaski, S.: Keyword Selection Method for Characterizing Text Document Maps. In: Proc. of ICANN99, Ninth International Conference on Artificial Neural Networks. Volume 1., London, IEEE (1999)
19. Kaski, S., Honkela, T., Lagus, K., Kohonen, T.: WEBSOM – Self-Organizing Maps of Document Collections. Neurocomputing **21** (1998)
20. Rauber, A.: LabelSOM: On the Labeling of Self-Organizing Maps. In: Proc. of the International Joint Conference on Neural Networks, IJCNN'99, Washington, DC (1999)
21. Whitman, B., Smaragdis, P.: Combining musical and cultural features for intelligent style detection. In: Proc. of the 3rd International Conference on Music Information Retrieval, Paris, France (2002)
22. Baumann, S.: Artificial Listening Systems: Modellierung und Approximation der subjektiven Perzeption von Musikähnlichkeit. PhD thesis, Technical University of Kaiserslautern (2005)
23. Knees, P., Pohle, T., Schedl, M., Widmer, G.: Combining Audio-based Similarity with Web-based Data to Accelerate Automatic Music Playlist Generation. In: Proc. of the 8th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR'06), Santa Barbara, California, USA (2006)
24. Whitman, B.: Learning the meaning of music. PhD thesis, Massachusetts Institute of Technology (2005)
25. Mandel, M., Ellis, D.: Song-Level Features and Support Vector Machines for Music Classification. In: Proc. of the 6th International Conference on Music Information Retrieval (ISMIR'05), London, UK (2005)
26. Aucouturier, J.J., Pachet, F.: Improving Timbre Similarity: How high is the sky? Journal of Negative Results in Speech and Audio Sciences **1**(1) (2004)
27. Pohle, T., Knees, P., Schedl, M., Widmer, G.: Automatically Adapting the Structure of Audio Similarity Spaces. In: Proc. of the 1st Workshop on Learning the Semantics of Audio Signals (LSAS'06), Athens, Greece (2006).