HOW MUCH METADATA DO WE NEED IN MUSIC RECOMMENDATION? A SUBJECTIVE EVALUATION USING PREFERENCE SETS

Dmitry Bogdanov and Perfecto Herrera

Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain

{dmitry.bogdanov,perfecto.herrera}@upf.edu

ABSTRACT

In this work we consider distance-based approaches to music recommendation, relying on an explicit set of music tracks provided by the user as evidence of his/her music preferences. Firstly, we propose a purely content-based approach, working on low-level (timbral, temporal, and tonal) and inferred high-level semantic descriptions of music. Secondly, we consider its simple refinement by adding a minimum amount of genre metadata. We compare the proposed approaches with one content-based and three metadata-based baselines. As such, we consider content-based approach working on inferred semantic descriptors, a tag-based recommender exploiting artist tags, a commercial black-box recommender partially employing collaborative filtering information, and a simple genre-based random recommender. We conduct a listening experiment with 19 participants. The obtained results reveal that although the low-level/semantic content-based approach does not achieve the performance of the baseline working exclusively on the inferred semantic descriptors, the proposed refinement provides significant improvement in the listeners' satisfaction comparable with metadata-based approaches, and surpasses these approaches by the number of novel relevant recommendations. We conclude that the proposed content-based approach refined by simple genre metadata is suited for music discovery not only in the long-tail but also within popular music items.

1. INTRODUCTION

Music recommendation is a challenging topic in the Music Information Research community. A rapid growth of digital music industry has led to vast amounts of music available for easy and fast access. Nevertheless, finding relevant and novel music is a difficult task for listeners, especially in the situation when new music appears every day. To fulfill their

needs, researchers and practitioners strive for better recommendation systems, which are able to facilitate music search and retrieval based on aggregated user profiles, or simple queries-by-example specified by users. To this end, improvements of suitable underlying user models and/or music similarity measures are necessary. Currently, the state-of-the-art approaches to music recommendation exploit both metadata information about music items (metadata-based approaches) and the information extracted from the audio signal itself (content-based approaches). Moreover, there exist hybrid approaches utilizing both types of information.

Possible metadata includes editorial information, social tags, and user listening/consumption behavior in form of listening statistics, such as playcounts and artist charts, sell histories, and user ratings. This information is found to be effective to provide satisfactory recommendations for users when dealing with popular music and operating on large collaborative filtering datasets. Nevertheless, the disadvantages of using metadata lie in the long-tail and cold-start problems [6]. A system may not have sufficient and correct metadata, including social tags, user ratings, or even editorial information, for unpopular items. This can significantly limit the quality of recommendations or even make them impossible. Moreover, gathering such metadata requires time and a large user base, which complicates the workability of the system on initial stages even for popular items.

In contrast, content-based information, extracted from the audio itself, can be valuable to overcome these problems as it can be used independently of the popularity of music items or availability of a user base. A number of research works exist on both content-based music similarity measures, or distances, [1] suitable for music recommendation, and approaches to user modeling. Objective content-based distances generally employ sets of low-level timbral, temporal, and tonal descriptors and/or high-level descriptors inferred from the low-level ones [2, 4, 5, 16, 17, 20]. Different works evidence usefulness of high-level semantic descriptions employed in place of, or in addition to, low-level music descriptions in the task of assessing music similarity [2, 4, 5]. There are also evidences that content-based

---

[1] We will refer to any music similarity measure with a term "distance".

approaches can be close, or even comparable, to successful metadata-based approaches in terms of the relevance of recommendations [1, 3, 15], especially in the long tail. In addition to objective distances, their personalization according to a concrete user is considered in some works. Alternatively, there exist research on user models for music recommendation which employ classification into interest categories using content-based information [8, 10] or hybrid sources [19], apply distances starting from a set of preferred items in a content-based vector space [3, 13], or propose more complex hybrid probabilistic approaches [12, 21].

One of the problems of existing research on music recommendation lies in a difficulty to conduct comprehensive subjective evaluations with real listeners. Up to our knowledge, few existing research works involve evaluations with real participants, and they are significantly limited by the number of participants [3, 10] or by the number of evaluated tracks per approach [1, 14], being in a trade-off situation.

In the present work, we consider music recommendation approaches which are based on sets of music tracks explicitly given by users as an evidence of their musical preferences (the henceforth called *"preference sets"*). We focus on content-based and hybrid approaches, striving for both relevance and novelty of recommendations. It is important to highlight the novelty aspect, as the existing metadata-based approaches working on collaborative filtering principles are known to have a drawback to produce recommendations already familiar to listeners [6]. We follow the research presented in [3, 9] in another peculiarity of this work, namely, using explicitly given preference examples. Such an explicit strategy was shown to capture the essence of users' musical preferences being suitable for preference visualization and distance-based music recommendation. Although requiring additional user effort to provide a list of preferred tracks, this strategy does not require any "adaptation" period, which is common to the cold-start prone systems gathering implicit user information. Starting from this strategy, we strive to improve distance-based approaches to music recommendation, working on content, evaluate them in comparison to metadata-based approaches on real listeners, and understand to what extent metadata is necessary to make a satisfactory music recommender.

We propose two distance-based approaches to music recommendation working on content-based and hybrid information (Section 2.1). Firstly, we consider a complex distance combining a set of low-level (timbral, temporal, and tonal) and inferred high-level semantic descriptors. This distance has been successfully evaluated in the task of objective music similarity [4], but it requires additional attention in the context of music recommendation. Secondly, we consider how a minimum amount of metadata can improve purely content-based recommendations, and propose a filtering approach relying on single, but sufficiently de-

scriptive, genre tags to refine recommendations. We evaluate these approaches against four baselines (Section 2.2). As such, we consider a content-based distance working on semantic descriptors, being a component of the proposed complex distance, and three approaches working purely on metadata. We employ a semantic tag-based approach, which operates on artist tags obtained from the *Last.fm* [2] service, and a state-of-the-art commercial recommender on the example of *iTunes Genius*, [3] which relies on a collaborative "wisdom of crowds". We also consider genre-based recommendations as the simplest metadata-based baseline. Characterization of subjects is presented in Section 3.1, while Section 3.2 explains the listening experiment instructions, stimuli and procedure. Section 3.3 presents and discusses the evaluation results, and we conclude with general observations and lessons learned from this study in Section 4.

2. STUDIED APPROACHES

To provide recommendations from our music collection (the henceforth called *music collection*), the approaches we consider here apply distance measures from a set of tracks, given by the user as evidence of his/her musical preferences (a *preference set*) to the tracks in the collection. In order to create such a preference set, the user is asked to gather a minimal set of music tracks, which he/she believes to be sufficient to grasp or convey his/her musical preferences, and submit them in audio format (e.g. mp3) or by editorial metadata sufficient to reliably identify and retrieve each track. The amount of required tracks is not specified being left to a decision of the user. We retrieve or clean the editorial metadata for all provided tracks by means of audio fingerprinting [4] to be able to use metadata-based approaches. As the source for recommendations, we employed a large in-house music collection, covering a wide range of genres, styles, and arrangements. This collection contains 68K music excerpts (30 sec.) by 16K artists with a maximum of 5 tracks per artist. For consistency, in our experiments we assume each of the recommendation approaches to output 15 tracks by different artists (1 track per artist) not being present among the artists in the user's preferences set. Therefore, each approach applies an artist filter.

2.1 Proposed Approaches

2.1.1 Semantic/Low-level Content-based Distance (C-SEMLL)

As our first proposed approach, we follow the ideas presented in [4] and employ a complex content-based distance,

---

[2] http://last.fm, all tags were obtained on March, 2011.
[3] http://www.apple.com/itunes/features/, all experiments were conducted using iTunes 10.1.1.4 on March, 2011.
[4] We used MusicBrainz service: http://musicbrainz.org/doc/MusicBrainz_Picard.

which is a weighted combination of three components:

- A Euclidean distance on a set of timbral, temporal, and tonal descriptors with a preliminary principle component analysis.
- A timbral distance based on the Kullback-Leibler divergence between single Gaussian models of MFCCs.
- A simple tempo distance, based on matches of BPM and onset rate values.
- A semantic distance, working on a set of high-level semantic descriptors (genres, musical culture, moods, instrumentation, rhythm, and tempo) inferred by support vector machines (SVMs) from low-level timbral, temporal, and tonal features.

The latter semantic distance has been previously evaluated in the similar context of music recommendation based on preference sets [3], and was shown to surpass common low-level timbral approaches. The interested reader is referred to the aforecited literature for further details about the descriptors used, the component distances, and their weighting.

We retrieve recommendations using this distance by the following procedure. For each track X in the user's preference set (a recommendation source), we apply the distance to retrieve the closest track C_X (a recommendation outcome candidate) from the music collection and form a triplet $(X, C_X, distance(X, C_X))$. We sort the triplets by the obtained distances, delete the duplicates of the recommendation sources (i.e. each track from the preference set produces only one recommendation outcome), and apply an artist filter. We return the recommendation outcome candidates from the top 15 triplets as recommendations. If it is impossible to produce 15 recommendations due to the small size of the preference set (less than 15 tracks) or the applied artist filter, we increase the amount of possible recommendation outcome candidates per recommendation source.

2.1.2 Semantic/Low-level Content-based Distance Refined By Genre Metadata (C-SEMLL+M-GENRE)

We consider the inclusion of metadata in purpose to refine the recommendations provided by content-based methods on the example of C-SEMLL. We strive to include the minimum amount of metadata, preferably being low-cost to gather and maintain, but however sufficiently descriptive for effective filtering. The experiments conducted in [3] point us to the fact, that simple genre/style tags can be a reasonable source of information to provide recommendations superior to the common low-level timbral music similarity based on MFCCs. Therefore, we propose a simple filtering to expand the C-SEMLL approach. We apply the same sorting procedure, but we solely consider the tracks of the same genre labels as possible recommendation outcomes. Moreover, we suppose that increasing the specificity of genre tags to certain amount (e.g. from "rock" to "prog rock") would in-

crease the quality of filtering.

To this end, we annotate the music collection and the user's preference set with genre tags. Such information can be obtained for the music collections by manual expert annotations, from social tagging services, or can be already available in the ID3 tags for audio files or in other metadata description formats generated on the music production stage. As a proof-of-concept, we opt for obtaining artist tags with the *Last.fm* API to simulate manual single-genre annotations of each track. *Last.fm* provides tag information for both artists and tracks. We opt for artist tags due to the fact that track tags tend to be more sparse, generally more difficult to obtain, and can be insufficient for the music retrieval in the long tail, and assign to the tracks the same tags that were assigned to the artists.

We analyze a set of possible tags suitable for the music collection. For each track, we select the *Last.fm* artist tags with the maximum weight (100.0) and add them to the pool of possible tags for genre annotation ("top-tags"). We then filter the pool deleting the tags with less than 100 occurrences (this threshold was selected in accordance with the top-tag histogram and the collection size) and blacklisting the tags which do not correspond to genres ("60s", "80s", "under 2000 listeners", "japanese", "spanish", etc.) We then revise the music collection to annotate each track with a single top-tag. For each track, we consider the candidates among its artist tags, selecting the tags with the maximum possible weight, which are also present in the top-tag pool. If there are several candidates (e.g. both "rock" and "prog rock" have weight 100.0 and are present in the top-tag pool), we select the top-tag, which is the least frequent in the pool. Thereafter, we annotate the tracks from the user's preference set in the same manner using the created pool. The idea behind this procedure is to select the most salient tags (top-tags) for the music collection, skip possible tag outliers, and annotate each track with the most specific of these top-tags keeping the maximum possible confidence level.

2.2 Baseline Approaches

2.2.1 Semantic Content-based Distance (C-SEM)

As our first baseline, we employ a content-based distance, working on a set of inferred high-level semantic descriptors, which was used as a component of the complex distance in the C-SEMLL approach (see Section 2.1.1). Using this distance, we retrieve recommendations with the same sorting procedure as followed for the C-SEMLL approach.

2.2.2 Artist Similarity based on Last.fm Tags (M-TAGS)

Alternatively, we consider a metadata-based distance working on the artist level. We gather social tags provided by the *Last.fm* API for the artists from the preference set and the music collection. For each artist, the API provides a

weight-normalized tag list with weights in the $[0, 100.0]$ interval. We select a minimum weight threshold of 10.0 to filter possibly inaccurate tags. We assign the resulting tags to each track in the preference set and the music collection. We then apply the latent semantic analysis [11, 18] to reduce the dimensionality to 300 latent dimensions. We apply the Pearson correlation distance [6] on the resulting topic space, and retrieve recommendations with the same procedure as followed for the C-SEMLL.

2.2.3 Black-box Similarity by iTunes Genius (M-GENIUS)

We consider commercial black-box recommendations obtained from the *iTunes Genius* playlist generation algorithm. Given a music collection and a query, this algorithm is capable to generate a playlist by means of the underlying music similarity measure, which works on metadata and partially employs collaborative filtering of large amounts of user data (music sales, listening history, and track ratings) [1]. From the preference set we randomly select 15 tracks annotated by artist, album, and track title information, sufficient to be recognized by *Genius*. For each of the selected tracks (a recommendation source), we generate a playlist, apply the artist filter, and select the top track as the recommendation outcome. We increase the amount of possible outcomes per source when it is impossible to produce 15 recommendations.

2.2.4 Random Tracks From the same Genre (M-GENRE)

Finally, as the simplest and low-cost metadata-based baseline, we consider random recommendations relying on genre categories of the user's preference set. We annotate the music collection and the user's preference set with genre labels by the same procedure as in the C-SEMLL+M-GENRE approach (see Section 2.1.2). We randomly preselect 15 tracks from the preference set and for each of the tracks we return a random track of the same genre label from the music collection. Again, we increase the amount of possible recommendation outcomes per recommendation source when it is impossible to produce 15 recommendations.

3. EVALUATION

3.1 Subjects

A total of 19 voluntary subjects (selected from the authors' colleagues, their acquaintances and families) were asked to provide their respective preference sets and additional information, including personal data (gender, age, interest for music, musical background), and a description of the strategy and criteria followed to select the music pieces. The participants were not informed about any further usage of the gathered data, such as giving music recommendations. The participants' age varied between 26 and 46 ($\mu = 33.72$, $\sigma = 4.65$). All participants showed a very high interest in music (rating with $\mu = 9.24$ and $\sigma = 1.01$, where 0 means no interest and 10 means passionate). In addition, 17 participants play at least one musical instrument. The number of tracks selected by the participants to convey their musical preferences was very varied, ranging from 10 to 178 music pieces ($\mu = 67.26$, $\sigma = 42.53$) with the median being 61 tracks. The time spent for this task also differed a lot, ranging from half an hour to 60 hours ($\mu = 6.22$, $\sigma = 15.06$) with the median being 2 hours. The strategy followed by the participants to gather preference sets varied as well. Driving criteria for the selection of tracks included musical genre, mood, uses of music (listening, dancing, singing, playing), expressivity, musical qualities, and chronological order. Taking into account this information, we expect our population to represent music enthusiasts.

3.2 Evaluation Methodology

We performed subjective listening tests on the 19 participants using our in-house music collection (see Section 2). One recommendation playlist per each of the 6 considered approaches was generated for each participant. Each playlist consisted of 15 tracks returned by the respected approach specifics. Due to the applied artist filter, the playlists neither contained more than one track of the same artist nor contained artists present in the preference set. We merged, randomized, and anonymized all playlists. This allowed to avoid any response bias due to presentation order, recommendation approach, or contextual recognition of tracks (e.g. by artist names) by participants. Moreover, the participants were not aware of the amount of recommendation approaches, their names and their rationales.

A questionnaire was given for the subjects to express different subjective impressions related to the recommended music. A "*familiarity*" rating ranged from the identification of artist and title (4) to absolute unfamiliarity (0), with intermediate steps for knowing the title (3), the artist (2), or just feeling familiar with the music (1). A "*liking*" rating measured the enjoyment of the presented music with 0 and 1 covering negative liking, 2 being a kind of neutral position, and 3 and 4 representing increasing liking for the musical excerpt. A rating of "*listening intentions*" measured preference, but in a more direct and behavioral way than the "*liking*" scale, as an intention is closer to action than just the abstraction of liking. Again this scale contained 2 positive and 2 negative steps plus a neutral one. Finally, an even more direct rating was included with the name "*give-me-more*" allowing just 1 or 0 to respectively indicate a request for, or a reject of, more music like the one presented. The users were also asked to provide title and artist for those tracks rated high in the familiarity scale. The textual meaning of the ratings was presented to the participants together with the rating values.

3.3 Evaluation Results

First, we manually corrected the familiarity rating when the artist/title, provided by the user, was wrong (hence a familiarity rating of "3" or, more frequently, "4", was sometimes lowered to 1). These corrections represented less than 3% of the total familiarity judgments.

Considering the subjective ratings used and our focus on music discovery, i.e. relevant and novel recommendations, we expect a good recommender system to provide high liking, listening intentions, and "give-me-more" ratings for a majority of the retrieved tracks and, most importantly, for low-familiarity tracks. We recoded user ratings for each evaluated track into 3 main categories - *hits*, *fails*, and *trusts* - referring to the type of the recommendation. In the case of liking, hits were the tracks which received low-familiarity rating (< 2) and a high (> 2) liking rating. Fails were the tracks having a low (< 3) liking rating. Trusts were the tracks which got a high familiarity (> 1) and a high (> 2) liking rating. We similarly recoded the intentions and "give-me-more" ratings, and obtained three different recommendation outcome categories per recommended track. We then combined the into a final category requiring the coincidence of all three outcome categories in order to consider it to be a hit, a fail, or a trust. Otherwise, the recommendation was considered as "unclear" (e.g. when a track is a hit using the liking, but it is a fail by other two indexes), which, in total, amounted to 20.4% of all recommendations. We excluded these recommendations from further analysis.

Table 1 reports the percent of each outcome category per recommendation approach. As we can see, the proposed C-SEMLL+M-GENRE approach yielded the largest amount of hits (32.0%), followed by M-TAGS (29.7%) and M-GENIUS (28.2%). The C-SEMLL+M-GENRE was the only (partially) content-based approach that provided considerably large amount of successful recommendations. We can evidence that inclusion of genre metadata improved the amount of hits by 11% for the C-SEMLL, making its refined version comparable to the metadata-based baselines. On the other side, the M-GENIUS and M-TAGS approaches provided the largest amount of trusts (18.3% and 10.6% respectively), while the rest of approaches yielded only scarce trusts (5.3% for C-SEMLL+M-GENRE, the rest below 3%). Trusts, provided their overall amount is low, can be useful for a user to feel that the recommender is understanding his/her preferences [1,7]. Nevertheless, their amount should not be excessive, especially in the use-case of music discovery. Finally, we can see that all recommendation approaches provided more than 33% of fails, which means that at least each third recommendation was possibly annoying for the user. In order to test if the approach and the outcome are associated (i.e. if certain approaches provide hit, fails or trust percents that are statistically different than those provided by other methods) we performed a chi-square test that

Approach	fail	hit	trust	unclear
C-SEMLL+M-GENRE	41.9	32.0	5.3	20.8
M-TAGS	38.9	29.7	10.6	20.8
M-GENIUS	33.1	28.2	18.3	20.4
M-GENRE	51.2	26.0	2.8	20.0
C-SEM	53.3	23.9	2.8	20.0
C-SEMLL	58.1	21.1	0.4	20.4

Table 1. Percent of fail, trust, hit, and unclear categories per recommendation approach.
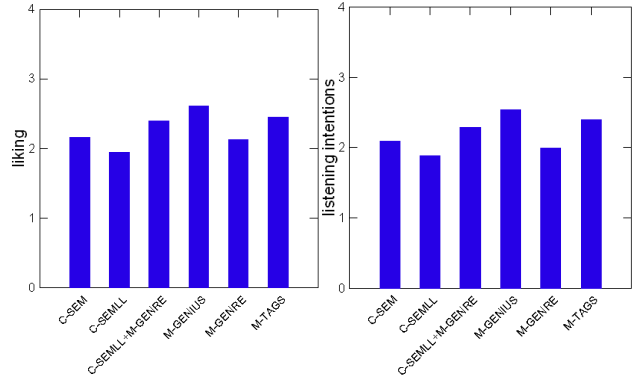

Figure 1. Means of liking and listening intentions ratings per recommendation approach.

provided support for that ($\chi^2(15) = 131.5$, $p < 0.001$).

In addition, we conducted three separate between-subjects ANOVAs in order to test the effects of the recommendation approaches on the liking, intentions, and "give-me-more" subjective ratings. The effect was confirmed in all of them ($F(5, 1705) = 15.237$, $p < 0.001$ for the liking rating, $F(5, 1705) = 14.578$, $p < 0.001$ for the intentions rating, and $F(5, 1705) = 11.420$, $p < 0.001$ for the "give-me-more" rating). Pairwise comparisons using Tukey's test revealed the same pattern of differences between the approaches, irrespective of the 3 tested indexes. It highlights the following groups with no statistically significant difference inside each group: 1) M-GENIUS, M-TAGS, and C-SEMLL+M-GENRE having the highest ratings, 2) C-SEM and C-SEMLL+M-GENRE, and 3) C-SEM, M-GENRE, and C-SEMLL having the lowest. Note, that these groups are partially intersected with the C-SEMLL+M-GENRE and C-SEM both belonging to two different groups. The mean liking and listening intentions ratings are presented in Figure 1.

4. CONCLUSIONS

We have considered different distance-based approaches to music recommendation, working on content information and metadata to generate recommendations from a set of music

tracks explicitly provided by a user as an evidence of her/his musical preferences. We proposed a complex content-based low-level/semantic approach and its simple refinement using genre labels as a minimum amount of metadata. We hypothesized that such single-genre information is considerably low-cost to gather and maintain meanwhile it is sufficiently descriptive for effective filtering.

The proposed approaches were evaluated against the four baselines on a population of 19 music enthusiasts. Considering purely content-based approaches, we did not find any improvements over the baseline semantic recommender using a complex low-level/semantic distance instead. This suggests that such a complex distance, previously found to overcome the semantic distance in the task of music similarity, is not well suited for the music recommendation use-case. Further study to reveal its nature will be necessary. Nevertheless, the refining of the proposed complex distance by simple genre labels showed a significant improvement. Furthermore, such a refined approach surpasses the considered metadata-based recommenders in terms of successful novel recommendations (hits) and provides satisfying recommendations, comparable to these baselines with no statistically significant difference.

The conducted evaluation corroborates a similar study presented in [3], in which similar patterns of no statistically significant difference between a content-based semantic distance and a simple genre-based baseline were found. The gap between both of them and commercial metadata-based recommendations, partially exploiting collaborative filtering data, was also shown there. We extend this results now with the proposed refining approach making possible to overcome such a gap. We may conclude that the proposed approach, operating on complex content-based distance, refined by simple genre metadata is well suited for the use-case of music discovery not only for the long-tail but also for popular items.

5. ACKNOWLEDGMENTS

6. REFERENCES

[1] L. Barrington, R. Oda, and G. Lanckriet. Smarter than genius? human evaluation of music recommender systems. In *Int. Society for Music Information Retrieval Conf. (ISMIR'09)*, pages 357–362, 2009.

[2] L. Barrington, D. Turnbull, D. Torres, and G. Lanckriet. Semantic similarity for music retrieval. In *Music Information Retrieval Evaluation Exchange (MIREX'07)*, 2007. http://www.music-ir.org/mirex/abstracts/2007/AS_barrington.pdf.

[3] D. Bogdanov, M. Haro, F. Fuhrmann, E. Gómez, and P. Herrera. Content-based music recommendation based on user preference ex-

amples. In *ACM Conf. on Recommender Systems. Workshop on Music Recommendation and Discovery (Womrad 2010)*, 2010.

[4] D. Bogdanov, J. Serrà, N. Wack, P. Herrera, and X. Serra. Unifying low-level and high-level music similarity measures. *IEEE Trans. on Multimedia*, 13(4):687–701, 2011.

[5] D. Bogdanov, J. Serrà, N. Wack, and P. Herrera. From low-level to high-level: Comparative study of music similarity measures. In *IEEE Int. Symp. on Multimedia (ISM'09)*, pages 453–458, 2009.

[6] O. Celma. *Music recommendation and discovery in the long tail*. PhD thesis, UPF, Barcelona, Spain, 2008.

[7] H. Cramer, V. Evers, S. Ramlal, M. Someren, L. Rutledge, N. Stash, L. Aroyo, and B. Wielinga. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, 18(5):455–496, 2008.

[8] M. Grimaldi and P. Cunningham. Experimenting with music taste prediction by user profiling. In *ACM SIGMM Int. Workshop on Multimedia Information Retrieval (MIR'04)*, pages 173–180, 2004.

[9] M. Haro, A. Xambó, F. Fuhrmann, D. Bogdanov, E. Gómez, and P. Herrera. The musical avatar - a visualization of musical preferences by means of audio content description. In *Audio Mostly (AM '10)*, 2010.

[10] K. Hoashi, K. Matsumoto, and N. Inoue. Personalization of user profiles for content-based music retrieval based on relevance feedback. In *ACM Int. Conf. on Multimedia (MULTIMEDIA'03)*, pages 110–119, 2003.

[11] M. Levy and M. Sandler. Learning latent semantic models for music from social tags. *Journal of New Music Research*, 37(2):137–150, 2008.

[12] Q. Li, S. H. Myaeng, and B. M. Kim. A probabilistic music recommender considering user opinions and audio features. *Information Processing & Management*, 43(2):473–487, 2007.

[13] B. Logan. Music recommendation from song sets. In *Int. Conf. on Music Information Retrieval (ISMIR'04)*, pages 425–428, 2004.

[14] C. Lu and V. S. Tseng. A novel method for personalized music recommendation. *Expert Systems with Applications*, 36(6):10035–10044, 2009.

[15] T. Magno and C. Sable. A comparison of signal-based music recommendation to genre labels, collaborative filtering, musicological analysis, human recommendation, and random baseline. In *Int. Conf. on Music Information Retrieval (ISMIR'08)*, pages 161–166, 2008.

[16] E. Pampalk. *Computational models of music similarity and their application in music information retrieval*. PhD thesis, Vienna University of Technology, 2006.

[17] T. Pohle, D. Schnitzer, M. Schedl, P. Knees, and G. Widmer. On rhythm and general music similarity. In *Int. Society for Music Information Retrieval Conf. (ISMIR'09)*, pages 525–530, 2009.

[18] M. Sordo, O. Celma, M. Blech, and E. Guaus. The quest for musical genres: Do the experts and the wisdom of crowds agree? In *Int. Conf. of Music Information Retrieval (ISMIR'08)*, pages 255–260, 2008.

[19] J. H. Su, H. H. Yeh, and V. S. Tseng. A novel music recommender by discovering preferable perceptual-patterns from music pieces. In *ACM Symp. on Applied Computing (SAC'10)*, pages 1924–1928, 2010.

[20] K. West and P. Lamere. A model-based approach to constructing music similarity functions. *EURASIP Journal on Advances in Signal Processing*, 2007:149–149, 2007.

[21] K. Yoshii, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno. Hybrid collaborative and content-based music recommendation using probabilistic model with latent user preferences. In *Int. Conf. on Music Information Retrieval (ISMIR'06)*, 2006.