THE QUEST FOR MUSICAL GENRES: DO THE EXPERTS AND THE WISDOM OF CROWDS AGREE?

Mohamed Sordo, Òscar Celma, Martín Blech, Enric Guaus
Music Technology Group
Universitat Pompeu Fabra
{msordo, ocelma, mblech, eguaus}@iua.upf.edu

ABSTRACT

This paper presents some findings around musical genres. The main goal is to analyse whether there is any agreement between a group of experts and a community, when defining a set of genres and their relationships. For this purpose, three different experiments are conducted using two datasets: the *MP3.com* expert taxonomy, and *last.fm* tags at artist level. The experimental results show a clear agreement for some components of the taxonomy (*Blues*, *Hip-Hop*), whilst in other cases (e.g. *Rock*) there is no correlations. Interestingly enough, the same results are found in the MIREX2007 results for audio genre classification task. Therefore, a multi–faceted approach for musical genre using expert based classifications, dynamic associations derived from the wisdom of crowds, and content–based analysis can improve genre classification, as well as other relevant MIR tasks such as music similarity or music recommendation.

1 INTRODUCTION

Music genres are connected to emotional, cultural and social aspects, and all of them influence our music understanding. The combination of these factors produce a personal organization of music which is, somehow, the basis for (human) musical genre classification. Indeed, musical genres have different meanings for different people, communities, and countries [2].

The use of musical genres has been deeply discussed by the MIR community. A good starting point is the review by McKay [5]. The authors suggested that musical genres are an inconsistent way to organize music. Yet, musical genres remain a very effective way to describe and tag artists.

Broadly speaking, there are two complementary approaches when defining a set of genre labels: (*i*) the definition of a controlled vocabulary by a group of experts or musicologists, and (*ii*) the collaborative effort of a community (social tagging). The goal of the former approach is the creation of a list of terms, organised in a hierarchy. A hierarchy includes the relationships among the terms; such as hyponymy. The latter method, social tagging, is a less formal bottom–up approach, where the set of terms emerge during the (manual) annotation process. The output of this approach is called folksonomy.

The aim of this paper is, then, to study the relationships between these two approaches. Concretely, we want to study whether the controlled vocabulary defined by a group of experts concord with the tag annotations of a large community.

Section 2 introduces the pros and cons of expert–based taxonomies and music folksonomies. To compare the similarities between both approaches, we gathered data from two different websites: a musical genre taxonomy from *MP3.com*, and a large dataset of artists' tags gathered from the *last.fm* community. Section 3 presents these datasets. The experimental results, presented in section 4, are conducted in order to analyse the relationships between the genres used in the *MP3.com* taxonomy, and the genre–tags annotated in the artist dataset from *last.fm*. Finally, section 5 concludes and summarizes the main findings.

2 MUSICAL GENRES CLASSIFICATION

2.1 Expert–based taxonomies

Depending on the application, taxonomies dealing with musical genres can be divided into different groups [6]: Music industry taxonomies, Internet taxonomies, and specific taxonomies.

Music industry taxonomies are created by recording companies and CD stores (e.g. RCA, Fnac, Virgin, etc.). The goal of these taxonomies is to guide the consumer to a specific CD or track in the shop. They usually use four different hierarchical levels: (1) Global music categories, (2) Sub-categories, (3) Artists (usually in alphabetical order), and (4) Album (if available).

Internet taxonomies are also created under commercial criteria. They are slightly different from the music industry taxonomies because of the multiple relationships that can be established between authors, albums, etc. The main property is that music is not exposed in a physical space (shelves). Obviously, exploiting the relationships among the items allows the end–user a richer navigation and personalization of the catalogue.

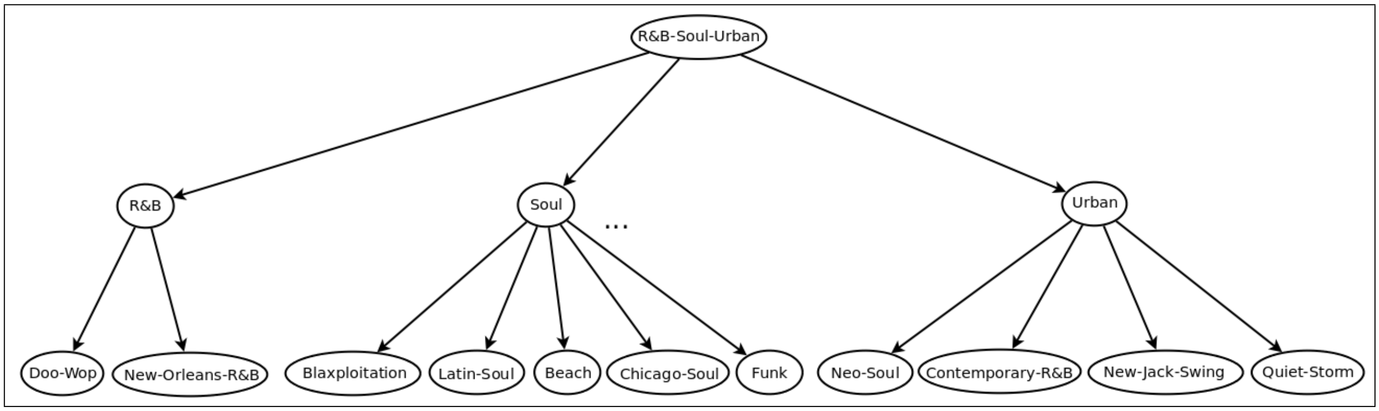

Figure 1. Partial view of the *MP3.com* taxonomy, starting with the seed genre *R&B–Soul–Urban*.

Furthermore, [6] shows that there is little consensus among the experts when defining a taxonomy. As an example, using three different musical genre taxonomies (*AllMusicGuide*, *Amazon*, and *MP3.com*) only 70 terms from more than 1500 were common in all the taxonomies.

2.2 Music Folksonomies

Since 2004, the explosion of Web 2.0 (e.g. tagging, blogging, user–generated content, etc.) questioned the usefulness of controlled vocabularies [11]. Internet sites with a strong social component, like *Last.fm*, allow users to tag music according to their own criteria. This scenario made the world of taxonomies even more complex.

Nowadays, users can organize their music collection using personal tags like *late night*, *while driving*, etc. As mentioned in the introduction, new strategies for music classification have emerged. Folksonomies exploit user–generated classification through a bottom–up approach [9].

On the one hand, this non-hierarchical approach allows users to organize their music with a better confidence. On the other hand, it creates difficulties for the design and maintenance of expert–based taxonomies, as new terms may emerge from time to time. Thus, in this scenario, up to date expert–based taxonomies become more and more difficult. Yet, it seems reasonable to analyse whether the genres derived from the tagging process share some patterns with the experts' controlled vocabulary.

3 DATASETS

3.1 Expert–based taxonomy from *MP3.com*

The *MP3.com* dataset was gathered during September 2005. Table 1 shows the relevant information about the genre taxonomy. Experts and musicologists from *MP3.com* identified 744 genres, and organized them in 13 different components, or in other words, the taxonomy has 13 seed-genres. The maximum depth is 6 levels, however, in most of the cases each component has 3 levels (plus the seed–genre at the top). Furthermore, this vocabulary still remains as it was three years ago (only a few more genres were added), showing its lack of evolution.

Total number of genres	744
Levels	7
Seed–genres	13
Num. genres at level 1	84
Num. genres at level 2	500
Num. genres at level 3	13
Num. genres at level 4	85
Num. genres at level 5	39
Num. genres at level 6	10

Table 1. Dataset gathered during September 2005 from the *MP3.com* expert–based taxonomy.

A partial view of the *MP3.com* taxonomy is depicted in Figure 1. It shows a component of the taxonomy. The seed genre is *R&B–Soul–Urban*, and the component consists of 3 different levels. A directed edge, e.g. *Urban* → *New-Jack-Swing*, represents the parent genre (*Urban*) and its subgenre(s), *New-Jack-Swing*.

3.2 Folksonomy from the *last.fm* community

A large dataset of artists' tags was gathered from the *last.fm* community during December 2007. Table 2 shows some basic information about the dataset. It is interesting to note that from the average number of tags per artist, 39% correspond to matched genres from the expert taxonomy, whilst the other 61% are distributed among other kinds of tags; including unmatched genres, decades, instruments, moods, locations, etc. For example, the artist *Jade* has the following tags (with their corresponding last.fm normalised weight):

```
Jade: urban(100), rnb(81), 90s(68),
      new jack swing(55), illinois(50),
      r and b(36), ...
```

Number of mapped genres	511
Number of artists	137,791
Number of distinct tags	90,078
Avg. tags per artist	11.95
Avg. *MP3.com* genres per artist	4.68

Table 2. Dataset gathered from the *last.fm* community during December 2007.

Nevertheless, since the experiments aim to analyse the agreement between expert genres and genre–tags, we need to match artists' tags with the expert defined genres.

3.2.1 Matching MP3.com genres and last.fm artist tags

In order to match those tags from the folksonomy that correspond to a genre in the expert taxonomy, a two-step process is followed:

- Compute a normalised form for all the folksonomy tags and expert genres, by:
  - converting them into lowercase,
  - unifying separators to a single common one,
  - treating some special characters (such as "&", which can be expanded to "and" and "n").
- Compute a string matching between the normalised folksonomy tags and expert genres.

The former step is inspired from [3]. For the latter, a string matching algorithm by Ratcliff and Metzener [8] is used to get all possible matches of a tag against a genre from the taxonomy. The similarity value goes from 0 to 1. Values close to 0 mean that the two strings are very dissimilar, and a 1 value means that the strings are identical. Deciding which is the threshold for identifying "nearly-identical" words is not trivial. Yet, [10] shows that a threshold of 0.85 gives the highest *F-measure*.

The following example shows artist *Jade*'s tags that are mapped to an *MP3.com* genre (*90s* and *illinois* tags disappear, and *rnb* and *r and b* are merged, combining their weights—with a maximum value of 100):

```
Jade: Urban(100), R&B(100),
      New-Jack-Swing(55)
```

Once the matching process is complete, the next step is to analyse whether the tagging behaviour of the community shares any resemblance with the expert taxonomy. The following section presents the experimental results.

4 EXPERIMENTAL RESULTS

In order to measure the agreement between expert genres and the genre–tags defined by the wisdom of crowds, we perform several experiments. Beforehand, we have to compute the similarities among genres. Section 4.1 explains the process of computing distances in the expert taxonomy (using the shortest path between two genres), and the tag distances in the folksonomy (by means of a classic Information Retrieval technique, called Latent Semantic Analysis).

The experiments are divided in two main groups. The first set of experiments deal with measuring the agreement at component level (a seed–genre and its subgenres). That is, to validate whether this taxonomy partition (13 components) correspond to the view of the community. Section 4.2 present these experiments. The other experiment focuses on the hierarchical structure (levels) of the expert taxonomy. In this experiment the goal is to reconstruct the taxonomy based on the genre distances from the folksonomy (section 4.3).

4.1 Computing genre distances

4.1.1 Expert taxonomy

To compute genre distances in the expert taxonomy we simply choose the shortest path between two genres, as an analogy with the number of mouse clicks to reach one genre from a given one (e.g. distance between genres *New–Jack–Swing* and *Soul* is 3, according to Figure 1). Since the taxonomy contains 13 seed genres, a virtual root node is added at the top, thus making the graph fully connected. This way we can compute the path between any genre in the graph. Whenever a path traverses the virtual root node, a penalty in the distance is added to emphasize that the two genres come from different components.

4.1.2 Folksonomy

Latent Semantic Analysis (LSA), plus cosine similarity, is used as a measure of distance among genres within the folksonomy. LSA assumes a latent semantic structure that lies underneath the randomness of word choice and spelling in "noisy" datasets [1], such as the one we are using. A significant paper that applies LSA in the music domain is [4]. The authors show the usefulness of social tags—in a low 10^2 space—to several relevant MIR problems, such as music similarity and mood analysis.

LSA makes use of algebraic techniques such as Singular Value Decomposition (SVD) to reduce the dimensionality of the Artist–Genres matrix. After this step, either artist or genre similarity can be computed using a cosine distance. Moreover, Information Retrieval literature [1, 7] states that, after raw data has been mapped into this *latent semantic space*, topic (in our case, genre) separability is improved.

	Folk	**Bluegrass**	**Country**	**Electronic-Dance**	**New-Age**	**Rock-Pop**	**Jazz**	**Hip-Hop**	**R&B-Soul-Urban**	**Gospel-Spiritual**	**Vocal-Easy-Listening**	**Blues**	**World-Reggae**
Folk	***0.184***	0.144*	0.048	0.002	0.040	-0.022	0.007	-0.005	0.003	0.095	0.001	-0.002	0.107
Bluegrass	0.144	***0.987***	0.738*	0.006	-0.018	0.008	-0.019	-0.006	-0.022	0.096	0.014	0.019	-0.033
Country	0.048	**0.738**	*0.430*	0.001	-0.034	0.048	0.007	0.007	0.009	0.054	0.031	0.008	-0.042
Electronic-Dance	0.002	0.006	0.001	*0.056*	0.019	0.012	0.025	0.004	0.019	0.036	**0.171**	0.002	-0.007
New-Age	0.040	-0.018	-0.034	0.019	***0.306***	0.125	0.001	0.022	0.018	0.068	0.102	0.037	0.303*
Rock-Pop	-0.022	0.008	0.048	0.012	0.125*	*0.043*	0.036	0.005	0.006	0.040	0.043	0.006	**0.132**
Jazz	0.007	-0.019	0.007	0.025	0.001	0.036	***0.211***	-0.008	0.030	-0.036	0.150	0.046	0.018
Hip-Hop	-0.005	-0.006	0.007	0.004	0.022	0.005	-0.008	***0.599***	-0.005	0.027	0.003	-0.005	0.021
R&B-Soul-Urban	0.003	-0.022	0.009	0.019	0.018	0.006	0.030	-0.005	***0.393***	0.355*	0.009	0.002	0.005
Gospel-Spiritual	0.095	0.096	0.054	0.036	0.068	0.040	-0.036	0.027	**0.355**	*0.134*	0.003	0.163	-0.015
Vocal-Easy-Listening	0.001	0.014	0.031	**0.171**	0.102	0.043	0.150	0.003	0.009	0.003	*0.167*	0.012	0.040
Blues	-0.002	0.019	0.008	0.002	0.037	0.006	0.046	-0.005	0.002	0.163	0.012	***0.657***	-0.004
World-Reggae	0.107	-0.033	-0.042	-0.007	**0.303**	0.132	0.018	0.021	0.005	-0.015	0.040	-0.004	*0.142*

Table 3. Confusion matrix for the inter–components coarse grained similarity. For clarification purposes, the diagonal contains the intra–component similarity. The values marked with an asterisk are as significative as the highest value (in bold).

For each artist we create a vector based on their (last.fm normalised) genre–tags' frequencies. Once the matrix is decomposed by columns, using SVD with 50 dimensions, we obtain genre similarities. For example, the closest genres to *Heavy Metal* in the semantic space are *Power Metal*, *British Metal* and *Speed Metal*, all with a similarity value above 0.6. On the other hand, similarity between *Heavy Metal* and *Pop* yields a near–zero value.

4.2 Agreement between expert and community genres

To measure the agreement between expert defined genre components and community genres we perform two experiments. The first one carries out a coarse–grained similarity (at genre component level), where the main goal is to *separate* the expert genre clusters according to the genre distances in the folksonomy. The second experiment performs a fine–grained similarity (at genre node level) in order to see the correlations between the genre distance in the taxonomy and the distance in the LSA space derived from the folksonomy.

4.2.1 Coarse-grained similarity

The first experiment aims to check how *separable* are the expert defined genre components according to the genre distances in the folksonomy (as defined in section 4.1.2). The experiment is performed in two steps: (*i*) compute the LSA cosine similarity among all the subgenres within a component (*intra–component* similarity); and (*ii*) compute the LSA cosine similarity among components, using the centroid of each component (*inter–component* similarity).

The results for intra–component similarity are presented in Figure 2. The most correlated components are *Bluegrass*, *Hip-Hop* and *Blues*. Note however that the *Bluegrass* component has only 3 subgenres mapped in our *last.fm* dataset. The components with less community–expert agreement are *Electronic-Dance* and *Rock-Pop*. For the latter genre, it is worth noting that it is an ill–defined seed-genre, and it is also the one including the highest number of subgenres. Some of these *Rock-Pop* subgenres are so eclectic that they could belong to more than one component. For instance, *Obscuro* subgenre is defined in *Allmusic* [1] as "*...a nebulous category that encompasses the weird, the puzzling, the ill-conceived, the unclassifiable, the musical territory you never dreamed*

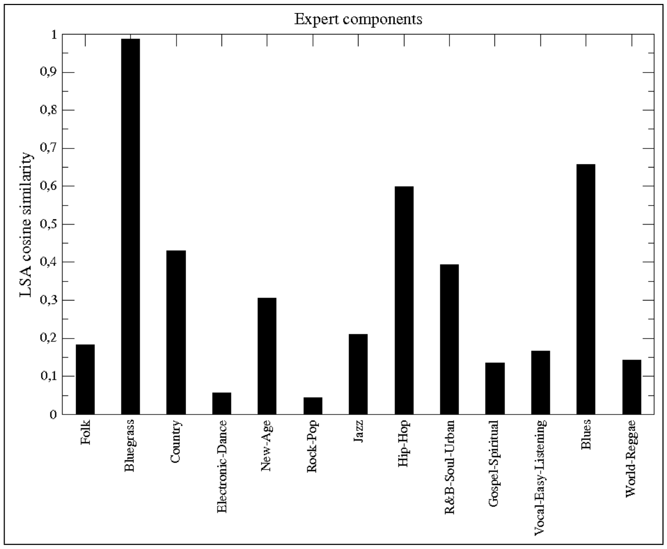

Figure 2. Intra–component coarse grained similarity.

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

existed".

Regarding the inter–component similarity, we proceed as follows: we compute the centroid vector of each component, and then compare it with the remaining components' centroids. The results are presented in table 3. Note that the results in the diagonal represent the intra–components similarity. For each row, we mark in bold the highest value. Subgenres of *Bluegrass*, *Hip-Hop* and *Blues*, as it has been observed for the intra–component case, are highly correlated in the semantic space. Thus, they are the ones with more agreement between the community and the experts classification. However, only *Hip-Hop* and *Blues* are clearly distinguishable from the rest. Furthermore, according to the community, *Bluegrass* and *Country* genres are very similar. Indeed, other available internet taxonomies, such as *Amazon* or *Allmusic*, include *Bluegrass* as a subgenre of *Country*. Similarly, *Gospel-Spiritual* genre is merged into *R&B-Soul-Urban*.

4.2.2 *Fine-grained similarity*

In this experiment we focus on the genre node level (instead of components). The hypothesis is that genres closer in the semantic space of the folksonomy should also be closer in the expert taxonomy, and vice versa. To validate this formulation a one–way Anova is performed. The independent groups are considered the path distances in the expert taxonomy (ranging from $1..10$, the diameter of the taxonomy), whilst the dependent variable is the LSA cosine distance.

Figure 3 depicts the box–and–whisker plot. Indeed, a large value of the *F–statistic* as well as a *p–value* $\ll 0.05$ corroborates the hypothesis. Furthermore, to determine the distances that are statistically significant we perform the Tukey's pairwise comparisons test. The results show that path distances 1 and 2 are significant among the rest of the distances, at 95% family–wise confidence level.

4.3 Reconstructing the taxonomy from the folksonomy

In this experiment we try to reconstruct the taxonomy from the folksonomy's inferred semantic structure. The reconstruction of the expert taxonomy from the folksonomy is based on the correct selection of a parent genre, according to the LSA cosine similarity derived from the folksonomy. We follow a bottom–up approach, starting from the leaves of each component. At each step of the process, we record the differences between the inferred and original taxonomies in order to have a similarity metric between them.

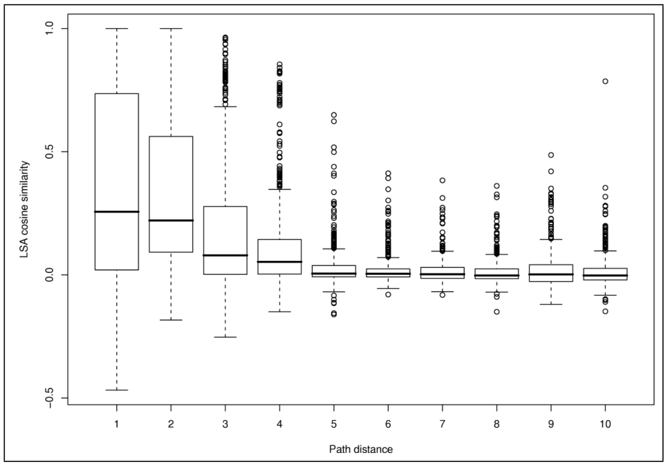

Figure 3. Box–and–whisker plot depicting the correlation between genre path distances in the taxonomy and semantic LSA cosine similarity. The Anova experiment ($p-value \ll 0.05$) shows that there is a statistical significance among path distances.

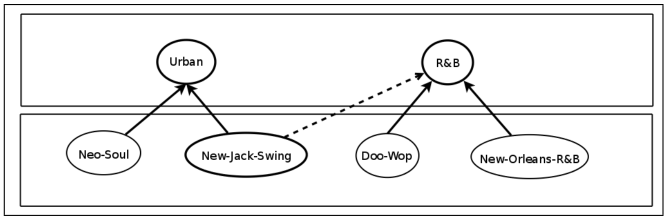

Figure 4. Reconstruction of the expert taxonomy from the folksonomy. Selection of a parent is made according to the LSA cosine similarity derived from the folksonomy.

The metrics used are: *mean reciprocal rank*, and *root hit*. The *mean reciprocal rank* (MRR) is a statistic widely used in Information Retrieval. The reciprocal rank (RR) is defined as the inverse of the correct answer's rank, $RR(tag) = 1/rank_{tag}$. For instance, given the *New–Jack–Swing* genre, see Figure 4, the closest genre parents (according to the LSA cosine distance) are: (1) *R&B*, (2) ***Urban***, (3) *Traditional–Gospel*, etc. The correct parent genre, *Urban*, is found in the second position, thus $RR(New - Jack - Swing) = \frac{1}{2} = 0.5$. Furthermore we compute whether the top–1 parent belongs to the same component as the child genre (**root hit**). In this example, it is a root hit because both the genre *New–Jack–Swing* and the selected (wrong) parent, *R&B*, belong to the same component, *R&B–Soul–Urban*.

Table 4 shows the results for each component. *Bluegrass*' perfect hit rate should be ignored, as the component has only 3 subgenres mapped. The lowest MRR is in *Rock–Pop* genre, which is also the highest (153 subgenres), and least cohesive component (lowest inter-genre similarity, see Table 3). *Hip–Hop*, on the other hand, is a highly cohesive component with a very high MRR. Finally, a lower MRR in *Folk* and *World–Reggae* could be interpreted as a consequence of the taxonomy being too geographically biased to English spoken territories. At the same time, disparate genres of all kinds, and from all over the world, are to be considered sub–genres of *Folk* and *World–Reggae*.

Component (size)	**Root Hits (%)**	**MRR**
Folk (22)	36.3	0.447
Bluegrass (3)	100	1.000
Country (35)	85.8	0.636
Electronic-Dance (36)	30.6	0.391
New-Age (5)	40.0	0.700
Rock-Pop (135)	48.2	0.295
Jazz (83)	83.2	0.638
Hip-Hop (22)	81.8	0.894
R&B-Soul-Urban (26)	75.4	0.694
Gospel-Spiritual (7)	38.6	0.558
Vocal-Easy-Listening (11)	27.3	0.446
Blues (43)	81.4	0.455
World-Reggae (83)	53.0	0.389
Weighted Avg. (511)	60.4	0.478

Table 4. Reconstruction of the expert–taxonomy, using genre similarity derived from the folksonomy.

5 CONCLUSIONS

This paper presented some interesting findings around musical genres. First of all, the consensus from a group of experts to create a universal taxonomy seems difficult. While expert taxonomies are useful for cataloguing and hierarchical browsing, the flat view of folksonomies allows better organization and access of a personal collection.

We presented three different experiments to analyse the agreement between expert–based controlled vocabulary and bottom–up folksonomies. The first two experiments focused on measuring the agreement between genres from the folksonomy and expert genres. A third experiment emphasized the hierarchical structure of a taxonomy, but using the information from a folksonomy. In all the experiments the conclusions were the same: some genres are clearly defined both from the experts and the wisdom of crowds, reaching a high agreement between these two views, while other genres are difficult to get a common consensus of its meaning.

All in all, experts, wisdom of crowds, and machines [2] agree in the classification and cohesion of some genres (e.g. *Blues*, *Hip-Hop*), and clearly disagree in others (*e.g. Rock*). A multi–faceted approach for musical genre using expert based classifications, dynamic associations derived from the community driven annotations, and content–based analysis would improve genre classification, as well as other relevant MIR tasks such as music similarity or music recommendation.

6 ACKNOWLEDGMENTS

This research has been partially supported by the e-Content plus project VARIAZIONI [3] .

7 REFERENCES

[1] J.R. Bellegarda. Latent semantic mapping. *Signal Processing Magazine, IEEE*, 22(5):70–80, Sept. 2005.

[2] F. Fabbri. A theory of musical genres: Two applications. *Popular Music Perspectives*, 1981.

[3] G. Geleijnse, M. Schedl, and P. Knees. The quest for ground truth in musical artist tagging in the social web era. In *Proceedings of the Eighth International Conference on Music Information Retrieval*, pages 525 – 530, Vienna, Austria, September 2007.

[4] M. Levy and M. Sandler. A semantic space for music derived from social tags. In *Proceedings of the Eighth International Conference on Music Information Retrieval*, Vienna, Austria, September 2007.

[5] C. Mckay and I. Fujinaga. Musical genre classification: Is it worth pursuing and how can it be improved? In *Proceedings of the Seventh International Conference on Music Information Retrieval*, Victoria, Canada, 2006.

[6] F. Pachet and F. Cazaly. A taxonomy of musical genres. In *Proc. Content-Based Multimedia Information Access*, 2000.

[7] C. H. Papadimitriou, H. Tamaki, P. Raghavan, and S. Vempala. Latent semantic indexing: A probabilistic analysis. In *Proocedings of the ACM Conference on Principles of Database Systems (PODS)*, pages 159–168, Seattle, 1998.

[8] JW Ratcliff and D. Metzener. Pattern matching: The Gestalt approach. *Dr. Dobb's Journal*, page 46, 1988.

[9] N. Scaringella, G. Zoia, and D. Mlynek. Automatic genre classification of music content: a survey. *Signal Processing Magazine, IEEE*, 23(2):133–141, 2006.

[10] A. Schuth, M. Marx, and M. de Rijke. Extracting the discussion structure in comments on news-articles. In *9th ACM International Workshop on Web Information and Data Management (WIDM 2007)*, pages 97–104, November 2007.

[11] C. Shirky. Ontology is overrated: Categories, Links, and Tags. *Online: http://shirky. com/writings/ontology_overrated. html*, 2005.

[2] See the results of MIREX 2007 in http://www.music-ir.org/mirex/2007/index.php/Audio_Genre_Classification_Results

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