MUSIC MOOD REPRESENTATIONS FROM SOCIAL TAGS

Cyril Laurier, Mohamed Sordo, Joan Serrà, Perfecto Herrera
Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
{cyril.laurier,mohamed.sordo,joan.serraj,perfecto.herrera}@upf.edu

ABSTRACT

This paper presents findings about mood representations. We aim to analyze how do people tag music by mood, to create representations based on this data and to study the agreement between experts and a large community. For this purpose, we create a semantic mood space from *last.fm* tags using Latent Semantic Analysis. With an unsupervised clustering approach, we derive from this space an ideal categorical representation. We compare our community based semantic space with expert representations from Hevner and the clusters from the MIREX Audio Mood Classification task. Using dimensional reduction with a Self-Organizing Map, we obtain a 2D representation that we compare with the dimensional model from Russell. We present as well a tree diagram of the mood tags obtained with a hierarchical clustering approach. All these results show a consistency between the community and the experts as well as some limitations of current expert models. This study demonstrates a particular relevancy of the basic emotions model with four mood clusters that can be summarized as: *happy*, *sad*, *angry* and *tender*. This outcome can help to create better ground truth and to provide more realistic mood classification algorithms. Furthermore, this method can be applied to other types of representations to build better computational models.

1. INTRODUCTION

Music classification by mood [1] recently emerged as a topic of interest in the Music Information Retrieval (MIR) community. The first task to tackle this problem is to find a relevant representation of mood. In this work, we study mood representations with a bottom-up approach, from a community point of view.

Several works have shown a potential to model mood in music (like [3–5] , see [6] for an extensive review). Although this task is quite complex, satisfying results can be achieved, especially if we concentrate on the mood expressed by the music rather than the mood induced [6]. However, almost every work differs in the way that it represents emotions. Similarly to psychological studies, there is no real agreement on a common model. Comparing these different techniques is a very arduous task. With the objective to evaluate several algorithms within the same framework, MIREX (Music Information Retrieval Evaluation eXchange) [7] organized a task on this topic for the first time in 2007. To do so, it was decided to frame the problem into a classification task with 5 mutually exclusive categories. However, it was shown that these clusters might not be optimal as we suspect some semantic overlap between categories [8]. In a nutshell, finding the right mood representation is complex.

In this study, we want to address this problem using data collected in an "everyday life" context (not in controlled laboratory settings like in psychological studies). From this data, we want to create a semantic space for mood. In [10], the authors studied the agreement between experts and a community (also based on *last.fm* tags) for genre classification. Levy in [11], studied how tags can be used for genre and artist similarity and proposed a visualization of certain words in an emotion space. Both studies inspired our approach of using social tags to compare the semantics of the wisdom of crowds with expert knowledge.

The goal of this paper is to create a semantic mood space where we can represent mood and compare it with existing representations. There are two main motivations for this study. First we aim to verify if the knowledge extracted from social tags and the knowledge from the experts (psychologists) converges. Then, we want to generate mood representations that can serve as a basis for further works like music mood classification. In Section 2, we expose the expert mood representations. In Section 3, we detail the dataset of tags and then, in Section 4, its transformation into a semantic space. In Section 5, we study the categorical representations. In Sections 6 and 7, we generate and analyze dimensional and hierarchical representations. Finally, Section 8 concludes and summarizes the main findings.

2. EXPERT REPRESENTATIONS

Two main types of representation coexist in the literature. The first one is the categorical model, using for instance basic emotions with around four or five categories including: *happiness*, *sadness*, *fear*, *anger* and *tenderness* [1]. Some works propose mood clusters like the eight clusters from Hevner [9] (see Figure 1) or the five clusters used in the MIREX Audio Mood Classification task, detailed

[1] In order to simplify the terminology, we will use the words emotion and mood independently for the same meaning: a particular feeling characterizing a state of mind

Clusters	Mood Adjectives
Cluster 1	passionate, rousing, confident, boisterous, rowdy
Cluster 2	rollicking, cheerful, fun, sweet, amiable/good natured
Cluster 3	literate, poignant, wistful, bittersweet, autumnal, brooding
Cluster 4	humorous, silly, campy, quirky, whimsical, witty, wry
Cluster 5	aggressive, fiery, tense/anxious, intense, volatile, visceral

Table 1. Clusters of mood adjectives used in the MIREX Audio Mood Classification task.

in Table 1. The second type of representation is the dimensional model, based originally on Russell's circumplex model of affect [2] (see Figure 2). The two dimensions mostly used are arousal and valence [2] .

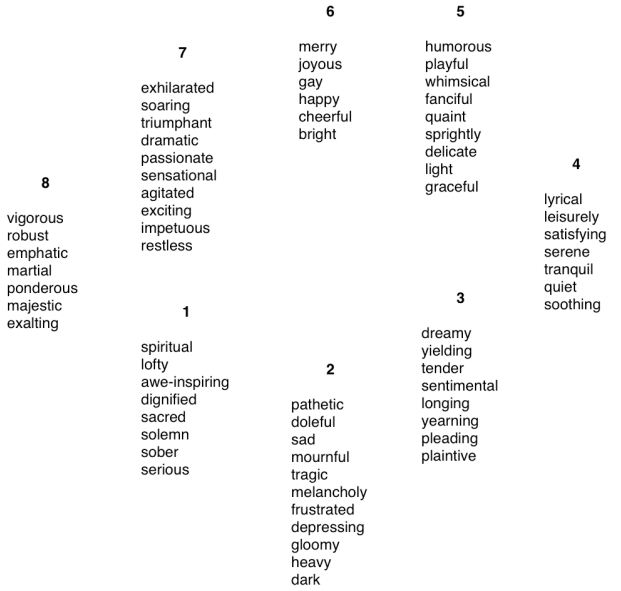

Figure 1. Hevner's [9] model with adjectives grouped into eight clusters.

3. DATASET

Our objective is to obtain a mood space based on social tags. In order to achieve this goal, we need two components: a list of mood words and social network data.

3.1 Mood list

For this study, we want to observe the way people use mood words in a social network. We selected words related to emotions based on the main articles in music and emotion research. We included words from different psychological studies like Hevner [9] or Russell [2]. We also added words representing basic emotions and other related adjectives [1]. Finally we aggregated the mood terms mostly used in MIR [6] and the ones selected for the MIREX task [8]. At the end of this process, we obtained a list of 120 mood words.

[2] In psychology, the term valence describes the attractiveness or aversiveness of an event, object or situation.

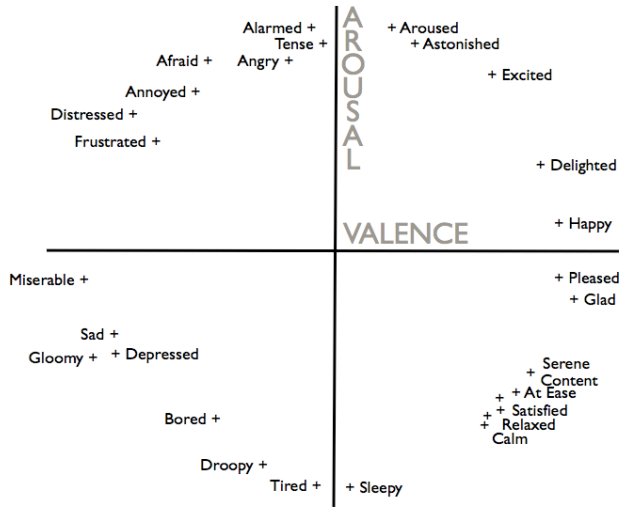

Figure 2. Russell's [2] circumplex model of affect with arousal and valence dimensions.

3.2 Social Tags

Last.fm [3] is a music recommendation website with a large community of users who are very active in associating tags with the music they listen to. With over 30 million users in more than 200 countries [4] , this social network is a good candidate to study how people tag their music. We crawled 6,814,068 tag annotations from 575,149 tracks in all main genres. From those, 492,634 tags were distinct. This huge dataset contains tags of any kind. From the original 120 mood words, 107 tags were found in our dataset. However some of them did not appear very often. We decided to keep only the tags that appeared at least 100 times, resulting in a list of 80 words. We also chose to keep the tracks were the same mood tag has been used by several users. This subset contains 61,080 tracks. We observe that the mood tags mostly used are *sad*, *fun*, *melancholy* and *happy*. For instance, the tag *sad* has been used 11,898 times in our dataset. On the contrary, the least used tags are *rollicking*, *solemn*, *rowdy* and *tense*, applied in less than 150 tracks. In average, a mood tag is used in 754 tracks.

4. SEMANTIC MOOD SPACE

We aim to compare mood terms by their co-occurences in tracks. Intuitively *happy* should co-occur more often with *fun* or *joy* than with *sad* or *depressed*. This co-occurence information included in the data we crawled from *last.fm* is embodied in a document-term matrix where the columns are track vectors representing tags.

The main problem we have when dealing with this matrix is its high dimensionality and its sparsity. Consequently, we applied a Latent Semantic Analysis (LSA) [12] to project the data into a space of a given lower dimensionality, while maintaining a good approximation of the distances between data points. This technique has been shown to be very efficient to capture tag representations for genre and artists

[3] http://www.last.fm
[4] http://blog.last.fm/2009/03/24/lastfm-radio-announcement

similarity [11]. LSA makes use of algebraic techniques such as Singular Values Decomposition (SVD) to reduce the dimensionality of the matrix. We decided to use a dimension of 100, which seems to be good trade-off for similarity tasks [11]. In the following experiments, we tried to change this dimension parameter (from 10 to 10 000 on a logarithmic scale), with no significant impact on the outcomes except less relevant results when selecting a too low or too high dimension. Once we have the data into this semantic space, we compute the distance between terms using the cosine distance. The distance values are included in the range $[0, 1]$. Here are some examples of distances between mood tags:

$d_{cos}(happy, sad) = 0.99$
$d_{cos}(cheerful, sleepy) = 0.97$
$d_{cos}(anger, aggressive) = 0.06$
$d_{cos}(calm, relaxed) = 0.03$

We observe that *happy* and *sad* are quite far from each other, as well as *cheerful* and *sleepy*. On the other hand, we note that *anger* is close to *aggressive* and that *calm* is similar to *relaxed*. Even if we show here some prototypical examples, values in the whole distance matrix intuitively make sense.

5. CATEGORICAL REPRESENTATIONS

To study the categorical mood representations, we first derive a folksonomy (community-based taxonomy) representation by means of unsupervised clustering from the social data. Then, we evaluate how the expert taxonomies fit into the semantic mood space.

5.1 Folksonomy representation

From our semantic space, we want to infer what would be the ideal categorical representation. To achieve this goal, we apply an unsupervised clustering method using the Expectation maximization (EM) algorithm. This algorithm and the implementation we used (WEKA) are described in [13]. The first important question to be answered is how many clusters should we consider. As we want this number to be inferred by the data itself, we used the *v-fold cross validation* algorithm. We divided the dataset in v folds, training on $v - 1$ folds and testing on the remaining one. We measure the log-likelihood computed for the observations in the testing samples. The results for the v replications are averaged to yield a single measure of the stability of the model. In Figure 3, we show the results of this process, displaying an average cost value (in our case 2 times the negative log-likelihood of the cross-validation data). Intuitively the lower is the value, the better is the cluster. To choose the "right" number of clusters, we look at the cost value while increasing the number of clusters. Practically, we stop when the mean cost value stops decreasing and select the current number of clusters.

We observe that the cost rapidly decreases with the number of clusters until four clusters. After that, it is stable and even increases, meaning that the data is overfitted. Consequently, the optimal number of clusters is four. Using this number for the EM algorithm, we obtained the clusters exposed in Table 2.

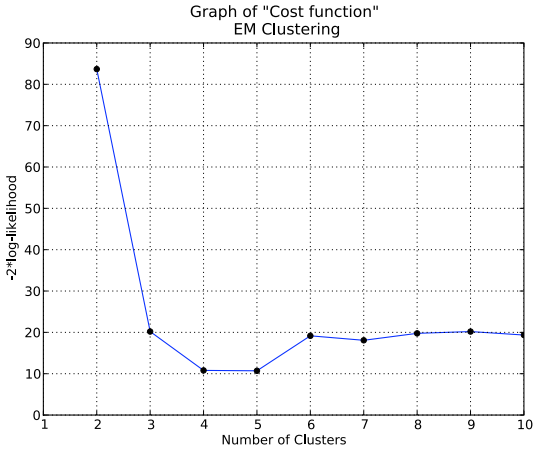

Figure 3. Plot of the cost values (2 times the negative log-likelihood) depending on the number of clusters.

Cluster 1	Cluster 2	Cluster 3	Cluster 4
angry	sad	tender	happy
aggressive	bittersweet	soothing	joyous
visceral	sentimental	sleepy	bright
rousing	tragic	tranquil	cheerful
intense	depressing	good natured	happiness
confident	sadness	quiet	humorous
anger	spooky	calm	gay
exciting	gloomy	serene	amiable
martial	sweet	relax	merry
tense	mysterious	dreamy	rollicking
anxious	mournful	delicate	campy
passionate	poignant	longing	light
quirky	lyrical	spiritual	silly
wry	miserable	wistful	boisterous
fiery	yearning	relaxed	fun

Table 2. Folksonomy representation. Clusters of mood tags obtained with the EM algorithm. For space and clarity reasons, we show only the first tags.

These four clusters are very similar to the categories posed by the main basic emotion theories [1]. Moreover, these clusters represents the four quadrants of the classical arousal-valence plane from Russell previously shown in Figure 2:

Cluster 1: angry (high arousal, low valence)
Cluster 2: sad, depressing (low valence, low arousal)
Cluster 3: tender, calm (high valence, low arousal)
Cluster 4: happy (high arousal, high valence)

To summarize, the semantic space we created is relevant and coherent with existing basic emotion approaches.

This result is very encouraging and assesses a certain quality of this semantic space. Moreover, it confirms that the community uses mood tags in a way that converges with the basic emotion theory from psychology.

5.2 Agreement between experts and community

In this section, we want to measure the agreement between experts and community representations. To do so, we performed a coarse-grained similarity, where we measured how *separable* the expert-defined mood clusters are in our semantic space. First, we computed the LSA cosine similarity among all moods within each cluster (intra-cluster similarity) and then we computed the dissimilarity among clusters, using the centroid of each cluster (inter-cluster dissimilarity). The expert representations we selected for this experiment are the eight clusters from Hevner (see Figure 1) where we could match more than 50% of the tags and the five clusters from the MIREX taxonomy (see Table 1) where all 31 tags were matched.

5.2.1 Intra-cluster similarity

For each cluster of the expert representations, we compute the mean cosine similarity between each mood tag in the cluster. The results for intra-cluster similarity are presented in Figure 4 for the Hevner representation and in Figure 5 for the MIREX clusters.

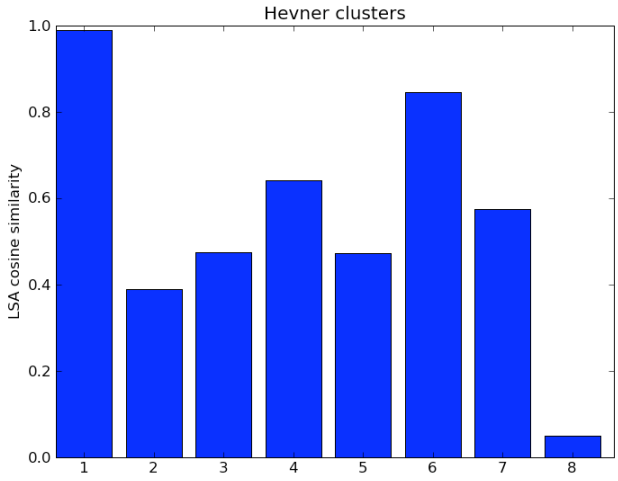

Figure 4. Intra-cluster cosine similarity for Hevner's representation.

In the results for the Hevner clusters, we note a high intra-cluster similarity value for cluster 1, which is the one including *spiritual* and *sacred* (please look at Figure 1 for the complete list). Cluster 6 performs also quite well (*joyous*, *bright*, *gay*, *cheerful*, *merry*). However we have poor intra-cluster similarity for cluster 8, which includes *vigorous*, *martial* and *majestic*. This might be because these words are also some of the less used in our dataset, but we hypothesize that they are less descriptive today than when the taxonomy was created (1936). Moreover, these words were selected for classical music which is not the main content of the *lasf.fm* music. The rest of the intra-cluster similarity values are in average quite low, meaning that this representation is not optimal in the semantic mood space.

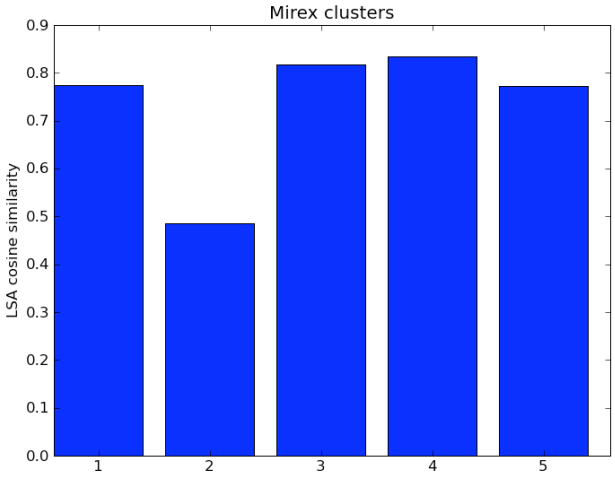

Figure 5. Intra-cluster cosine similarity for MIREX representation.

For the MIREX clusters, we remark that the lowest intra-cluster similarity is for cluster 2 (*sweet*, *good natured*, *cheerful*, *rollicking*, *amiable*, *fun*). Maybe is it quite clear that this category is about *happy* music, however the words used are not so common and may lower this value. In average, the intra-cluster similarity value is quite high for this representation. For comparison purpose, we note that the intra-cluster similarity of the folksonomy representation has an average intra-cluster similarity value of 0.82 (see Table 4). Obviously, as the folksonomy representation was made from the semantic space itself, it has better results than the other models.

In this part, we have looked at the consistency inside each cluster, however it is also crucial to look at the distances between clusters to evaluate the quality of the clustering representations.

5.2.2 Inter-cluster dissimilarity

To measure how *separable* are the different clusters, we compute the mean cosine distance from each cluster centroid to the other cluster centroids. If we look at our folksonomy representation clusters from Section 5.1, the cosine distance between centroids of clusters are all quite high (0.9 in average, see Table 4). This is not very suprising as the representation was designed with this data.

In Table 3, we show the confusion matrix of the inter-cluster dissimilarity for the MIREX clusters. We notice that the lowest value is between cluster 1 and cluster 5, meaning that these clusters are quite similar. This finding correlates with the results from the MIREX task, in which the confusion between these two classes was found significant [8]. However the confusion between clusters 2 and 4, also relevant in the MIREX results analysis, is not reflected here. Additionally, we observe that the most separated clusters (5 and 2), are also the less confusing in the MIREX results. Looking at the confusion matrix for the

	C1	C2	C3	C4	C5
C1	0	0.74	0.128	0.204	0.108*
C2	0.74	0	0.859	0.816	**0.876**
C3	0.128	0.859	0	0.319	0.265
C4	0.204	0.816	0.319	0	0.526
C5	0.108*	**0.876**	0.265	0.526	0

Table 3. Confusion matrix for the inter-cluster dissimilarity for the MIREX clusters (C1 means cluster 1, C2 cluster 2 and so on). The values marked with an asterisk are the most similar and in bold are the less similar values.

Hevner clusters (not shown here for space reasons), we remark that the highest values (dissimilarity above 0.95) are between clusters 7 and 8, and between clusters 1 and 2. On the contrary, the lowest value (0.09) is between clusters 1 and 4. Indeed both clusters have words than can appear similar like *spiritual* and *serene* for instance. We summarize the results of both intra and inter-cluster measures for the different taxonomies in Table 4.

Mood Taxonomy	Intra-cluster similarity	Inter-cluster dissimilarity
Hevner	0.55	0.70
MIREX	0.73	0.56
Folksonomy	0.82	0.9

Table 4. Intra-cluster similarity and inter-cluster dissimilarity means for each mood taxonomy.

In a nutshell, the Hevner clusters are less consistent but are more separated than the MIREX ones. Indeed, even if the latter has more intra-cluster similarity, it suffers from confusions between some categories as reflected in our results.

6. DIMENSIONAL REPRESENTATION

Dimensional representation is an important paradigm in emotion studies. To project our semantic mood space into a bi-dimensional space, we used the Self-Organizing Map algorithm (SOM). We decided to use SOM for its topology properties and because it stresses more on the local similarities and distinguishes groups within the data. Because less than half of the Russell's adjectives are present in our dataset, we prefer to compare qualitatively more that quantitatively the expert and the community models. We trained a SOM and mapped each tag onto its best-matching unit in the trained SOM. In Figure 6, we plot the resulting organization of mood tags (for clarity reasons, we show here a subset of 58 tags).

We observe in the 2D projection four main parts. At the top-left, terms related to *aggressive*, below *calm* and other similar words, at the top-right tags related to *sad* and below words close to *happiness*. We notice the four clusters corresponding to the basic emotions and our folksonomy representation mentioned in Section 5.1. This is somehow

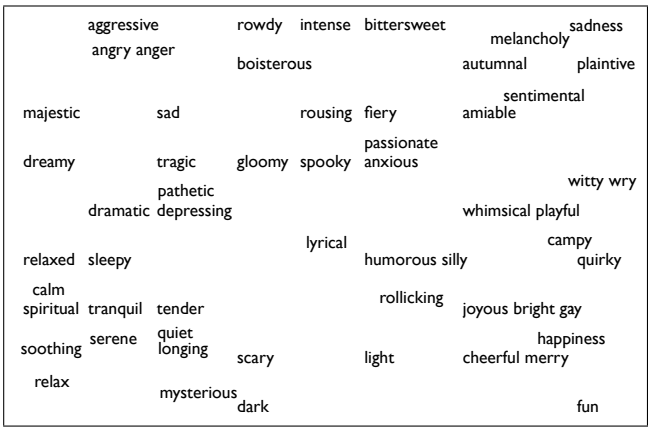

Figure 6. Self-Organizing Map of the mood tags in the semantic space.

expected as we already got these clusters from this data. However, having the same results with a second technique confirms our findings. Comparing with Russell's dimensions, we find that the diagonal from top-left to bottom-right is of high arousal. On the contrary, the diagonal from top-right to bottom-left is of low arousal. The vertical axis represents the valence dimension. Even though the 2D representation is not equal, there is a clear correlation between the community and the experts when framing the problem into two dimensions.

7. HIERARCHICAL REPRESENTATION

The semantic mood space can be visualized in many different ways. In this part we experimented hierarchical clustering techniques to produce a tree diagram (dendrogram). We applied a common agglomerative hierarchical clustering method with a complete linkage [14] and the cosine metric. We used the hcluster [5] implementation. With the 20 most used tags in our dataset, we computed the clustering and plot the resulting dendrogram in Figure 7 .

Although there exists some dendrogram representation of emotions in the psychology literature [1], the comparison is complex because many of the terms employed are not present in our dataset and also because finding the right metric to measure the similarity between both is not trivial. The hierarchical clustering starts with two branches. Looking at the tags of this first branching, we observe a very clear separation in arousal with *dreamy* and *calm* on the left and *angry* and *happy* on the right. Then the two following branching (resulting in four clusters) represents the four basic emotions also found as the best categorical representation in Section 5 (in order in the dendrogram: *calm*, *sad*, *angry* and *happy*). This confirms another time our findings about the relevancy of these four clusters. Moreover, we notice that the first separation is related to arousal, often considered as the most important dimension. The remaining branches group together similar terms like *angry* and *aggressive* or *sad* and *depressing*.

[5] http://code.google.com/p/scipy- cluster

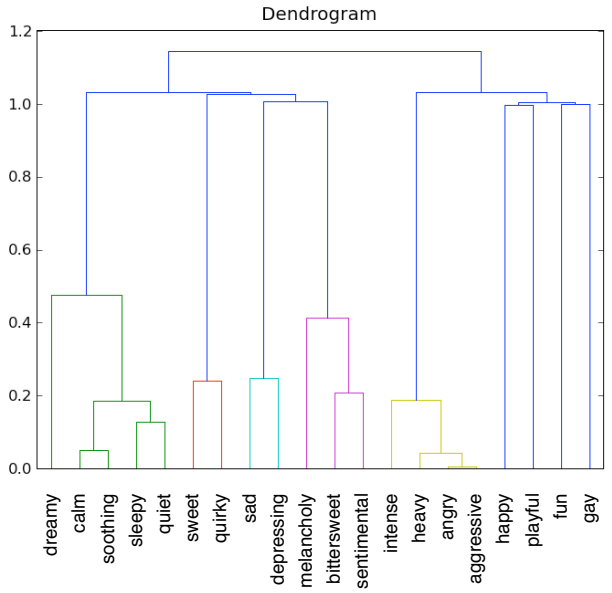

Figure 7. Dendrogram of the 20 most used tags.

8. CONCLUSIONS

This paper presented convergent evidence about mood representations. We created a semantic mood space based on a community of users from *last.fm*. We derived different representations from this data and compared them to the expert representations. We demonstrated that the basic emotions: *happy*, *sad*, *angry* and *tender*, are very relevant to the social network. We also found that the arousal and valence dimensions are pertinent. Moreover we have shown that both Hevner's and MIREX representations have advantages and limitations when evaluated in the semantic mood space. The former having better separated clusters and the latter having more consistent clusters. Observations on the confusion and similarity between MIREX clusters confirmed results from previous analysis. We also presented a dendrogram visualization validating again the basic emotion point of view and offering a new representation of the mood space. All these findings show the relevancy of using a mood semantic space derived from social tags. Folksonomy representations can be used in tasks like mood classification or regression to improve the quality of the audio content processing algorithms. We can also imagine a visualization of a user emotional states based on his listening habits or history. Moreover, one's musical library can be mapped and explored with a folksonomy representation derived from the whole social network or a particular subset. Finally this approach can be generalized to find other domain-specific representations.

9. ACKNOWLEDGMENTS

We want to thank the people from the Music Technology Group (Universitat Pompeu Fabra, Barcelona). Data from this work is available at: *http://mtg.upf.edu/people/claurier* This research has been partially funded by the EU Project PHAROS IST-2006-045035.

10. REFERENCES

[1] P. N. Juslin and J. A. Sloboda: *Music and Emotion: Theory and Research*, Oxford University Press, 2001.

[2] J. A. Russell: "A circumplex model of affect," *Journal of Personality and Social Psychology*, No. 39, pp. 1161, 1980.

[3] T. Li and M. Ogihara: "Detecting emotion in music," *Proceedings of ISMIR, Baltimore, MD, USA*, pp. 239–240, 2003.

[4] Y. H. Yang, Y. C. Lin, Y. F. Su, and H. H. Chen: "A regression approach to music emotion recognition," *IEEE Transactions on audio, speech, and language processing*, Vol. 16, No. 2, pp. 448–457, 2008.

[5] C. Laurier, O. Meyers, J. Serrà, M. Blech, P. Herrera: "Music Mood Annotator Design and Integration," *7th International Workshop on Content-Based Multimedia Indexing, Chania, Crete*, 2009.

[6] C. Laurier, P. Herrera: "Automatic Detection of Emotion in Music: Interaction with Emotionally Sensitive Machines," *Handbook of Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence*,Chap. 2, pp. 9–32, IGI Global, 2009.

[7] J. S. Downie: "The music information retrieval evaluation exchange (2005-2007): A window into music information retrieval research," *Acoustical Science and Technology*, Vol. 29, No. 4, pp. 247–255, 2008.

[8] X. Hu, S. J. Downie, C. Laurier, M. Bay, and A. F. Ehmann: "The 2007 MIREX audio mood classification task: Lessons learned," *Proceedings of ISMIR, Philadelphia, PA, USA*, pp. 462–467, 2008.

[9] K. Hevner: "Experimental studies of the elements of expression im music," *The American Journal of Psychology*, Vol. 48, No. 2, pp. 246–268, 1936.

[10] M. Sordo, O. Celma, M. Blech, and E. Guaus: "The Quest for Musical Genres: Do the Experts and the Wisdom of Crowds Agree?," *Proceedings of ISMIR, Philadelphia, PA, USA*, pp. 255–260, 2008.

[11] M. Levy and M. Sandler: "A Semantic Space for Music Derived from Social Tags," *Proceedings of ISMIR, Vienna, Austria*, 2007.

[12] S. Deerwester, S. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman : "Indexing by latent semantic analysis.," *Journal of the Society for Information Science*, Vol. 14, pp. 391–407, 1990.

[13] I. H. Witten and E. Frank: *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, 1999.

[14] R. Xu and D. C. Wunsch: *Clustering*, IEEE Press, 2009