EXTENDING THE FOLKSONOMIES OF FREESOUND.ORG USING CONTENT-BASED AUDIO ANALYSIS

Elena Martínez, Òscar Celma, Mohamed Sordo, Bram de Jong, Xavier Serra
Music Technology Group
Universitat Pompeu Fabra, Barcelona, Spain
elena.martinez@openbravo.com, ocelma@bmat.com, mohamed.sordo@upf.edu, bdejong@iua.upf.edu, xavier.serra@upf.edu

ABSTRACT

This paper presents an in–depth study of the social tagging mechanisms used in *Freesound.org*, an online community where users share and browse audio files by means of tags and content–based audio similarity search. We performed two analyses of the sound collection. The first one is related with how the users tag the sounds, and we could detect some well–known problems that occur in collaborative tagging systems (i.e. polysemy, synonymy, and the scarcity of the existing annotations). Moreover, we show that more than 10% of the collection were scarcely annotated with only one or two tags per sound, thus frustrating the retrieval task. In this sense, the second analysis focuses on enhancing the semantic annotations of these sounds, by means of content–based audio similarity (autotagging). In order to "autotag" the sounds, we use a k–NN classifier that selects the available tags from the most similar sounds. Human assessment is performed in order to evaluate the perceived quality of the candidate tags. The results show that, in 77% of the sounds used, the annotations have been correctly extended with the proposed tags derived from audio similarity.

1 INTRODUCTION

Since 2004, collaborative tagging seems a natural way for annotating objects, in contrast to using predefined taxonomies and controlled vocabularies. Internet sites with a strong social component (e.g. *last.fm*, *flickr*, and *del.icio.us*), allow users to tag web objects according to their own criteria. The tagging process can improve then, content organization, navigation, search and retrieval tasks [9].

Nowadays, in the multimedia domain, *prosumers* hold an important role. The term comes from producing and consuming at the same time: they create and annotate a vast amount of data. In fact, audiovisual assets can be manually and automatically described. On the one hand, users can organize their music collection using personal tags like: *late night*, *while driving*, *love*. On the other hand, content–based (CB) audio annotation can propose, with some confidence degree, audio related tags such as: *pop*, *acoustic guitar*, or *female voice*. It is clear that both approaches create a rich tag cloud representing the actual content. Still, automatic annotation based solely on CB cannot bridge the Semantic Gap. Hybrid approaches, exploiting both the wisdom of crowds and automatic content description, are needed in order to close the gap. In this sense, *Freesound.org*, a collaborative sound database, contains both elements: it allows users to annotate sounds, and they can also browse similar sounds to a given one, according to audio similarity. However, there are some sounds that are scarcely annotated, thus frustrating their retrieval using keyword–based search.

The main goal of this paper is to enhance semantic annotations in the *Freesound.org* sound collection, by means of content–based audio similarity. We propose an approach to "autotag" sounds based on the tags available in their most similar sounds.

SMC 2009, July 23-25, Porto, Portugal

2 COLLABORATIVE TAGGING

One of the most interesting aspects of collaborative tagging is that the whole community benefits from sharing information [17]. However, "collective tagging has also the potential to aggravate the problems associated with the fuzziness of linguistic and cognitive boundaries" [7]. Users' contributions produce a huge classification system that consists in an idiosyncratically personal categorization. Some of the main problems concerning collaborative tagging are: polysemy, synonymy and data scarcity. Furthermore, spelling errors, plurals and parts of speech also clearly affect a tagging system.

Sometimes, polysemous tags can return undesireable results. For example, in a music collection if one is searching using the tag *love*, the results can contain both love songs, and songs that users like it very much (i.e. a user that loves a *death metal Swedish* song, not related with the love theme).

Tag synonymy is also an interesting problem. Even though it enriches the vocabulary, it presents also inconsistencies among the terms used in the annotation process. For example, *bass drum* sounds can be annotated with the *kick drum* tag; but these sounds will not be returned when searching for *bass drum*. To avoid this problem, sometimes users tend to add redundant tags to facilitate the retrieval (e.g. using *synth*, *synthesis*, and *synthetic* for a given sound excerpt). Yet, there are some approaches to measure semantic relatedness between tags [3]. These metrics could be used to decrease the size of the vocabulary, and also for (automatic) query expansion to increase the recall in the sound retrieval task.

Finally, the scarcity and inequality nature of a collaborative annotation process—where usually a few sounds are well annotated, and the rest contain very few tags—limits the coverage retrieval of a collection.

3 RELATED WORK

In [16], the authors propose a query–by–semantic audio information retrieval system. The proposed system can learn the relationships between acoustic information and words (tags) from a manually annotated audio collection. The learning task is based on a supervised multiclass labeling model, with a multinomial distributions of words over a predefined vocabulary.

Torres et. al propose a method to construct a musically meaningful vocabulary [15]. By means of acoustic correlation using canonical component analysis (sparse CCA), they can remove from the vocabulary those noisy words (not related with the actual audio content) that have been inconsistently used by human annotators.

The *bag–of–frames* (BOF) approach has been extensively used to describe timbrical properties of an audio signal. This approach is used to extract mid–level descriptions from music signals, such as their genre or instrument, but it is also used to perform timbre similarity between songs. In [1], the authors find out that this approach tends to generate false positives songs which are irrelevantly close to many other songs. These songs are called hubs, and the authors propose measures to quantify the “hubness” of a given song. This property affects any system that uses timbrical features to compute content–based audio similarity.

Cano has studied the strengths and limitations of audio fingerprinting, and suggests that it can be extended to allow content–based similarity search, such as finding similar sounds using query–by–example [2]. Similarly to our approach, [14] proposes a non–parametric strategy for automatically tagging songs, using content–based audio similarity to propagate tags from annotated songs to similar, non–annotated, songs.

In [5], the authors present a method to recommend tags to unlabeled songs. Automatic tags are computed by means of a set of boosted classifiers (Adaboost), in order to provide tags to tracks poorly (or not) annotated. This method allows music recommenders to include in a playlist unheard music that otherwise would be missed, enhancing the novelty component of the recommendations.

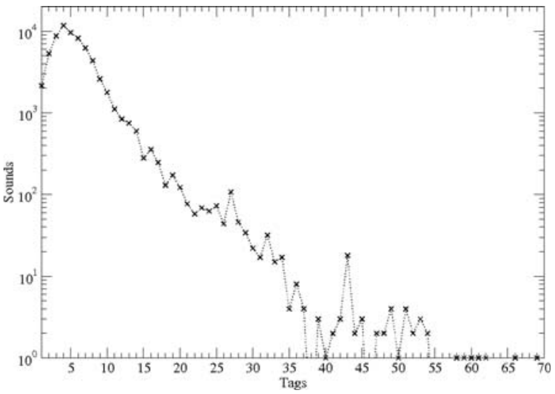

Figure 1. A linear–log plot depicting the number of tags per sound. Most of the sounds are annotated using 3–5 tags, and only a few sounds are annotated with more than 40 tags.

4 THE FREESOUND.ORG COLLECTION

Freesound.org is a collaborative sound database where people from different disciplines share recorded sounds and samples under the Creative Commons license, since 2005. The initial goal was to giving support to sound researchers, who often have trouble finding large sound databases to test their algorithms. After four years since its inception, *Freesound.org* serves more than 23,000 unique visits per day. Also, there is an engaged community—with almost a million registered users—accessing more than 66,000 uploaded sounds.

Yet, only few dozens of users uploaded hundreds of sounds, whilst the rest uploaded just a few. In fact, 80% of the users uploaded less than 20 sounds, and only 8 users uploaded more than one thousand sounds each. It is worth noting that these few users can highly influence the overall sound annotation process.

4.1 Tag behaviour

In this section we provide some insights about the tag behaviour and user activity in the *Freesound.org* community. We are interested in analyzing how users tag sounds assets, as well as the concepts used when tagging. The data, collected during March 2009, consists of around 66,000 sounds annotated with 18,500 different tags

Figure 1 shows the number of tags used to annotate the audio samples. The x-axis represent the number of tags used per sound. We can see that most of the sounds are annotated using 3–5 tags. Also, around 7,500 sounds are insufficiently annotated using only 1 or 2 tags. These sounds represent more than 10% of the whole collection. It would be desirable, then, to—automatically—recommend relevant tags to these scarcely annotated sounds, enhancing their descriptions. This is the main goal of the experiments presented in section 5.

Interestingly enough, in [2], the author analyzed a sound effects database, which was annotated by only one expert. A similar histogram distribution to the one presented in Figure 1 was obtained. Specifically, most of the sounds were annotated by the expert using 4 or 5 tags, as it is our case. This could be due to human memory constraints when assigning words to sounds or to any object, in order to describe them [11]. Based on Figure 1, we classify the sounds in three different categories, according to the number of tags used. Table 1 shows the data for each class.

Table 1. Sound–tag classes and the number of sounds in each category.

	Tags per sound	Sounds
Class I	1–2	7,481
Class II	3–8	42,757
Class III	> 8	7,148

Tag frequency distribution is presented in Figure 2. The x-axis refers to the 18,500 tags used, ranked by descending frequency. On the one hand, 44% of the tags were applied only once. This reflects the subjectivity of the tag process. Thus, retrieving these sounds in the heavy tail area is nearly impossible using only tag–based search (to overcome this problem, *Freesound.org* offers a content–based audio similarity search to retrieve similar sound samples). On the other hand, just 27 tags were used to annotate almost the 70% of the whole collection. The best fit of the tag distribution is obtained with a log–normal function, $\frac{1}{x}e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}$, with parameters mean of log $\mu = 1.15$, and standard deviation of log, $\sigma = 1.46$ [4].

The top–5 most frequent tags are presented in Table 2, and it gives an idea about the nature of the sounds available in the *Freesound.org* collection. Field–recording is the most frequent tag used to describe 6,787 different sounds. All these frequent tags are very informative when describing the sounds, in contrast to the photo domain in *flickr.com*, were popular tags are considered too generic to be "useful" [13].

Table 2. Top–5 most frequent tags from Figure 2.

Rank	Tag	Frequency
1	field–recording	6,787
2	noise	5,650
3	loop	5,487
4	electronic	4,329
5	synth	4,307

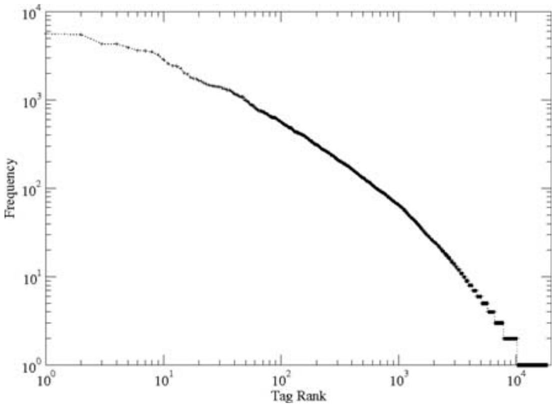

Figure 2. A log–log plot showing the tag distribution in *Freesound.org*. The curve follows a log–normal distribution, with mean of log $\mu = 1.15$, and standard deviation of log, $\sigma = 1.46$.

4.2 Tag categorization

In order to understand the vocabulary that the *Freesound.org* community uses when tagging sounds, we mapped the 18,500 different tags to broad categories (hypernyms) in the Wordnet [1] semantic lexicon. In some cases, a given tag matches multiple entries, so we bound the tag (noun or verb) to the highest ranked category. The selected Wordnet categories are: *(i)* artefact or object, *(ii)* organism, being, *(iii)* action or event, *(iv)* location, and *(v)* attribute or relation. Yet, 20.3% of the tags remain unclassified.

Most of the tags (38%) are related with objects (e.g. *seatbelt*, *printer*, *missile*, *guitar*, *snare*, etc.), or about the qualities and attributes of the objects (30%); such as state attributes (*analog*, *glitch*, *scratch*), or magnitude relation characteristics (*bpm*). Then, some tags (19%) are classified as an action (*hiss*, *laugh*, *glissando*, *scream*, etc.), whilst 11% are related with organisms (*cat*, *brass band*, etc.). Finally, only a few tags (2%) were bound to locations (e.g. *iraq*, *vietnam*, *us*, *san francisco*, *avenue*, *pub*, etc.). Therefore, we conclude that the tags are mostly used to describe the objects that produce the sound, and the characteristics of the sound. In this case, the wisdom of crowds concords with the studies of [12] and [6]. The former study focused on the attributes of the sound itself without referencing the source causing it (e.g *pitchiness*, *brightness*), while the latter introduced a taxonomy of sounds, on the assertion that they are produced by means of interaction of materials.

[1] `http://wordnet.princeton.edu/`

5 EXPERIMENTS

Our goal is to evaluate the quality of the recommended tags, for some specific sounds available in *Freesound.org*. By means of content–based audio similarity, our algorithm selects a set of candidate tags for a given sound (autotagging process). Then, the evaluation process is based on human assessment. Three subjects validated each candidate tag for all the sounds in the test dataset.

5.1 Dataset

The sounds selected for the experiments were a subset of the Class I (see Table 1). We selected those sounds whose tags' frequency was very low (i.e. rare tags, in the ranking of $\sim 10^4$ in Figure 2). In fact, all the sounds which were annotated with one tag whose frequency was equal to 1 were selected. Also, for the sounds annotated with 2 tags, we selected those which had at least one tag with frequency 1. The test dataset for the experiments consists of 260 sounds. The goal here is to automatically extend the annotation of these sounds, unsufficiently annotated with one or two very rare tags.

5.2 Nearest–neighbor classifier

We used a nearest neighbor classifier (k–NN, $k = 10$) to select the tags from the most similar sounds of a given sound. The choice of a memory–based nearest neighbor classifier avoids the design and training of every possible tag. Another advantage of using an NN classifier is that it does not need to be redesigned nor trained whenever a new class of sounds is added to the system. The NN classifier needs a database of labeled instances and a similarity distance to compare them. An unknown sample will borrow the metadata associated with the most similar registered sample.

Based on the results from [2], the similarity measure used is a normalized Manhattan distance of audio features belonging to three different groups: a first group gathering spectral and temporal descriptors included in the MPEG-7 standard [10]; a second one built on Bark Bands perceptual division of the acoustic spectrum, using the mean and variance of relative energies for each band; and, finally a third one, composed of Mel-Frequency Cepstral Coefficients (20) and their corresponding variances [8]. The normalized Manhattan distance of the above enumerated features is:

$$d(x,y) = \sum_{k=1}^{N} \frac{|x_k - y_k|}{(max_k - min_k)} \tag{1}$$

where x and y are the vectors of audio features, N the dimensionality of the feature space, and max_k and min_k the maximum and minimum values of the *k–th* feature.

5.3 Procedure

Our technique for calculating the candidate tags consists on finding the *10–th* most similar sounds from the *Freesound.org* database, for a given seed sound of the test dataset. That is, given a seed sound, we get the tags from the similar sounds. A tag is proposed as a candidate if it appears among the neighbors over a specific threshold. For example, a threshold of 0.3, means that a tag is selected as candidate when it appears at least in 3 sounds of the 10 nearest neighbors. This way we select the set of candidate tags for each sound in the test dataset.

The experiments have been computed using two thresholds: 0.3 and 0.4. When using a threshold of 0.3 the number of candidate tags is higher than for 0.4, but also there are more "noisy" or potentially irrelevant tags, since it is using a less constrained approach. Afterwards, all the candidate tags will be evaluated by human assessment. The differences between both thresholds is presented in section 6.1.

5.4 Evaluation

In order to validate the candidate tags for the test sounds, we use human assessment. The aim is to evaluate the perceived quality of the candidate tags. It is worth noting that neither Precision nor Recall measures are applicable as the test sound contains only two or less tags, and these are very rare in the vocabulary. We performed a listening experiment where the subjects were asked to listen to the sounds, and decide whether they agreed or not with the candidate tags. For each candidate tag, they had to select one of these options: *Agree* (recommend candidate tag), *Disagree* (do not recommend), or *Don't know*. Each sound was rated by three different subjects.

Similar to [16], to evaluate the results we group human responses for each sound s, and score them in order to compact them into a single vector per sound. The length of the vector is the number of candidate tags of s. Each value of the vector, w_{s,t_i}, contains the weight of the subjects' scores for a candidate tag t_i in sound s. If a subject agrees with the candidate tag, the score is $+1$, -1 if disagrees, and 0 if she does not know. The formula for calculating the weight of the candidate tag in s is:

$$\mathbf{w}_{s,t_i} = \frac{\#(PositiveVotes) - \#(NegativeVotes)}{\#Subjects} \tag{2}$$

A candidate tag is recommended to the original sound if $\mathbf{w}_{s,t_i}$ is greater than zero, otherwise, the tag is rejected (either because it is a bad recommendation, or the subjects cannot judge the quality of the tag). For example, given a candidate tag t_i for s, if the three subjects scored, respectively, $+1$, -1, $+1$ (two of them agree, and one disagree), the final weight is $\mathbf{w}_{s,t_i} = 1/3$. Since this value is greater than zero, t_i is considered a good tag to be recommended.

Furthermore, we use $\mathbf{w}_{s,t_i}$ to compute the confidence agreement among the subjects. First, we consider all the sounds where the system proposed j candidate tags, S_j. We sum, for each sound $s \in S_j$, the weights of all the candidate tags t_i whose values were greater than zero. Then, we divide this value with the total score that the candidate tags would had if all the subjects would agree. The formula for calculating the agreement of S_j sounds, A_j, is:

$$A_j = \frac{\sum_{s \in S_j} \left[\mathbf{w}_{s,t_i} > 0\right]}{\#Subjects \cdot \left[\sum_{s \in S_j} length\left(s\right)\right]} \tag{3}$$

Similarly, to compute the agreement of the bad candidate tags, we use the weights of candidate tags whose values were lesser than zero ($\mathbf{w}_{s,t_i} < 0$), in the numerator of the equation 3. Finally, to get the total agreement for all the sounds in the test set, A_{total}, we use the weighted mean of all A_j, according to the number of sounds in A_j.

6 RESULTS

6.1 Perceived quality of the recommended tags

Using 10–NN and the content–based audio similarity, and setting a threshold of 0.3, the system proposed a total of 781 candidate tags, distributed among the 260 sounds of the test dataset. Besides that, setting a threshold of 0.4 the system proposes 358 candidate tags, which represents almost the half compared with a threshold of 0.3.

Table 3 shows the human assessment results. As expected, a slightly higher percentage of candidate tags were recommended with a threshold of 0.4 (66.23%). Yet, using a threshold of 0.3, more than half of the candidate tags (56.6%) were finally recommended to the original sounds, with an agreement confidence of 0.74. This human agreement is sufficiently high to rely on the perceived quality of the recommended tags. The rest of the candidate tags (43.4%) were not recommended, either because the tags recommended were not appropiated (31.59%), or the tags were not sufficiently informative (11.41%). Even though with a threshold of 0.3 we get less percentage of recommended tags, the absolute number of candidate tags is more than twice the ones with a threshold of 0.4. Therefore, we can consider a threshold of 0.3 a good choice for this task.

6.2 Recommended tags per class

On the one hand, using a threshold of 0.4 we are able to enhance the annotation of half of the sounds (128 sounds out of 260). On the other hand, with a threshold of 0.3, we have enhanced the annotation of 200 sounds, which represent the 77% of the sounds in the test dataset used. The rest of the sounds (60) from the test set did not get any plausible tags to extend its current annotation.

Table 3. Percentage of recommended tags, with confidence agreement among the subjects. The table shows the results using thresholds 0.3 and 0.4 (in parenthesis, it is shown the total number of candidate tags).

Threshold	Recommend tag	%	A_{total}
0.3 (781)	Yes	56.60%	0.74
	No	31.59%	0.62
	Don't know	11.41%	—
0.4 (358)	Yes	66.23%	0.78
	No	23.11%	0.58
	Don't know	10.66%	—

Table 4. Number of sounds in each category, after automatically extending the annotations of 200 sounds from the test dataset.

	Tags per sound	Sounds
Class I	1–2	20
Class II	3–8	171
Class III	> 8	9

Table 4 shows the results using a threshold of 0.3, and it classifies the 200 autotagged sounds according to the classes defined in Table 1. Originally, all the test sounds belonged to Class I. We can observe now the number of sounds per class, after extending the annotation of these 200 sounds. Note that most of the sounds have 3 or more tags (Class II), and some even have more than 8 tags (Class III). However, there are 20 sounds still belonging to Class I. This happens because before the experiment they only had one tag, and now they have another one, the one recommended.

The results obtained so far look promising; using a simple classifier we were able to automatically extend sound annotations that were difficult to retrieve. Furthermore, due to the classifier method used (k–NN), there is a strong correlation among the more frequently proposed tags, and their frequency of usage (rank position in Figure 2). The ten most proposed tags are also in the top–15 ranking of frequency use. Although our approach is prone to popular tags, once the sounds are autotagged it allows the users to get a higher recall of those scarcely annotated sounds when doing a keyword–based search.

7 CONCLUSIONS

This paper presents an analysis of the *Freesound.org* collaborative database, where the users share and browse sounds by means of tags, and content–based audio similarity search. First we studied how users annotate the sounds in the database, and detected some well–known problems in collaborative tagging, such as polysemy, synonymy, and the scarcity of the existing annotations.

Regarding the experiments, we selected a subset of the sounds that are rarely tagged, and proposed a content–based audio similarity to automatically extend these annotations (autotagging). Since the sounds in the test set contained only one or two rare tags, neither precision nor recall were applicable, so we used human assessment to evaluate the results. The reported results show that 77% of the test collection were enhanced using the recommended tags, with a high agreement among the subjects.

As future work, we are planning to extend the experiments using more sounds. In this case, automatic evaluation is needed. A possible solution is to select sounds belonging to similar sound categories (e.g all the percussive sounds scarcely annotated), and follow the same procedure of finding similar sounds from the *Freesound.org* database. So, the recommended tags should also belong to the same sound category. We are also working on a hybrid approach that combines tag similarity and content–based similarity to improve the recommendations of the similar sounds.

8 ACKNOWLEDGMENTS

The authors would like to thank the Freesound community for making Freesound.org a great research platform on which to study many interesting sound problems. We also want to acknowledge the help of many researchers of our research group in the technical aspects of this paper.

9 REFERENCES

[1] J.-J. Aucouturier and F. Pachet. A scale-free distribution of false positives for a large class of audio similarity measures. *Pattern Recognition*, 41(1):272–284, 2008.

[2] P. Cano. *Content-Based Audio Search from Fingerprinting to Semantic Audio Retrieval*. PhD thesis, Universitat Pompeu Fabra, 2007.

[3] C. Cattuto, D. Benz, A. Hotho, and G. Stumme. Semantic analysis of tag similarity measures in collaborative tagging systems. In *Proceedings of the 3rd Workshop on Ontology Learning and Population*, Patras, Greece, July 2008.

[4] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Reviews*, June 2007.

[5] D. Eck, P. Lamere, T. Bertin-Mahieux, and S. Green. Automatic generation of social tags for music recommendation. In *Advances in Neural Information Processing Systems 20*, pages 385–392. MIT Press, Cambridge, MA, 2008.

[6] W. W. Gaver. What in the world do we hear? an ecological approach to auditory event perception. *Ecological Psychology*, 5(1):1–29, 1993.

[7] S. Golder and B. A. Huberman. The structure of collaborative tagging systems, 2005.

[8] P. Herrera, A. Yeterian, and F. Gouyon. Automatic classification of drum sounds: A comparison of feature selection methods and classification techniques. In *Proceedings of the Second International Conference on Music and Artificial Intelligence*, pages 69–80, London, UK, 2002.

[9] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. In *Proceedings of the 3rd European Semantic Web Conference*, pages 411–426, Budva, Montenegro, June 2006. Springer.

[10] B. Manjunath, P. Salembier, and T. Sikora. *Introduction to MPEG 7: Multimedia Content Description Language*. Ed. Wiley, 2002.

[11] G. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information, 1956.

[12] P. Schaeffer. *Trait des objects musicaux*. Editions du Seuil, Paris, 1966.

[13] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *Proceeding of the 17th international conference on World Wide Web*, pages 327–336, New York, NY, USA, 2008. ACM.

[14] M. Sordo, C. Laurier, and O. Celma. Annotating music collections: How content-based similarity helps to propagate labels. In *Proceedings of 8th International Conference on Music Information Retrieval*, Vienna, Austria, 2007.

[15] D. Torres, D. Turnbull, L. Barrington, and G. Lanckriet. Identifying words that are musically meaningful. In *Proceedings of 8th International Conference on Music Information Retrieval*, Vienna, Austria, 2007.

[16] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet. Towards musical query-by-semantic-description using the cal500 data set. In *Proceedings of 30th International SIGIR Conference*, pages 439–446, New York, NY, USA, 2007. ACM.

[17] J. Walker. Feral hypertext: when hypertext literature escapes control. In *Proceedings of the 16th conference on Hypertext and hypermedia*, pages 46–53, New York, NY, USA, 2005. ACM.