Class-based tag recommendation and user-based evaluation in online audio clip sharing

Frederic Font[a], Joan Serrà[b], Xavier Serra[a]

[a]*Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain*
[b]*Artificial Intelligence Research Institute (IIIA-CSIC), Spanish National Research Council, Bellaterra, Spain*

Abstract

Online sharing platforms often rely on collaborative tagging systems for annotating content. In this way, users themselves annotate and describe the shared contents using textual labels, commonly called tags. These annotations typically suffer from a number of issues such as tag scarcity or ambiguous labelling. Hence, to minimise some of these issues, tag recommendation systems can be employed to suggest potentially relevant tags during the annotation process. In this work, we present a tag recommendation system and evaluate it in the context of an online platform for audio clip sharing. By exploiting domain-specific knowledge, the system we present is able to classify an audio clip among a number of predefined audio classes and to produce specific tag recommendations for the different classes. We perform an in-depth user-based evaluation of the recommendation method along with two baselines and a former version that we described in previous work. This user-based evaluation is further complemented with a prediction-based evaluation following standard information retrieval methodologies. Results show that the proposed tag recommendation method brings a statistically significant improvement over the previous method and the baselines. In addition, we report a number of findings based on the detailed analysis of user feedback provided during the evaluation process. The considered methods, when applied to real-world collaborative tagging systems, should serve the purpose of consolidating the tagging vocabulary and improving the quality of content annotations.

Keywords: Tag recommendation, User study, Folksonomy, Freesound

1. Introduction

Free-form semantically-meaningful textual labels, called tags, are extensively used in online sharing platforms for describing and annotating contents. Systems that provide the functionality for making these annotations are normally referred to as collaborative tagging systems. Several problems arise when users annotate shared and/or online resources [9]. The most typical ones are tag scarcity, the use of different tags to refer to a single concept (synonymy), the ambiguity in the meaning of certain tags (polysemy), the commonness of typographical errors, the use of user-specific naming conventions, or the use of different languages. To minimise some of these problems, tag recommendation systems can be employed to suggest potentially relevant tags during the annotation process [14]. As users are exposed to the suggestions of the system, the annotation process partially shifts from the creation of textual labels to the recognition of tags in a list [23], and thus all users receive a certain common influence from the system. Hence, tag recommendation serves the purpose of consolidating the vocabulary of collaborative tagging systems [13].

In general, tag recommendations are either based on content analysis of online resources or in the other tags that users introduce during the annotation process. In the case of content-based recommendations, a typical approach consists in, given a resource to be described, defining a neighbourhood of other resources (based on some similarity measure) and then recommending tags that are used to annotate resources in this neighbourhood [12, 24]. Another approach is the use of machine learning techniques to learn mappings between tags and content features [15, 25, 26]. On the other side, there are tag recommendation strategies which are based on the tags that users introduce during the annotation process itself, prior to the moment of the recommendation. Disadvantages of these strategies compared to content-based recommendation methods are that they require the existence of at least one tag to provide recommendations, whereas content-based recommendation systems can provide recommendations to resources with no associated tags or other metadata. Nevertheless, tag recommendation methods based on the tags that users introduce during the annotation process have the advantage of not requiring any specific processing of the content of the resources being annotated, thus being typically less expensive in terms of computation resources and being more easily generalisable to other multimedia domains. These methods usually consider the *folksonomy* (i.e., the set of associations between tags, users and content resources) of a collaborative tagging system to estimate tag similarity from their resource co-occurrence. In this way, candidate tags can be selected according to their similarity to the introduced tags, and a sorting algorithm can rank them in terms of estimated relevance [4, 8, 14, 22]. In previous work, we described and evaluated a general scheme for folksonomy-based tag recommendation in collaborative tagging systems [7]. Out of that scheme, eight particular methods were proposed which form the basis of the method presented in this work.

Besides content-based and folksonomy-based tag recommendation systems, other approaches have been described in the literature. Anderson et al. [1] describe a tag recommendation system for Flickr[1], a well known photo sharing site, which combines both content-based recommendations (by training a predictive model that learns the mapping between tags and extracted content image features) with folksonomy-based recommendations (following an strategy very similar to [22]). Naaman and Nair[19] describe another tag recommen-

[1] `www.flickr.com`

dation system for Flickr, which takes advantage of the geolocation metadata attached to images and recommends tags that other users employed in close areas. Chen et al. [3] describe a tag recommendation system for video resources which crawls the web for information about these videos and identifies keywords to recommend as tags.

Although it is quite common to personalise tag recommendation systems to the tagging behaviour of particular users by promoting, for example, tags that users introduced in past annotations [2, 8, 14, 16, 18, 20], most of the current systems do not introduce direct user feedback in the evaluation loop. Thus recommendations are generally evaluated using traditional information retrieval approaches based on the comparison of tag rankings produced by different methods, or using precision and recall metrics computed after a tag prediction task [2, 7, 8, 16, 18, 20]. To the best of our knowledge, only three studies perform some kind of user-based evaluation. Sigurbjörnsson and Zwol [22] automatically generate tag recommendations for several images from a Flickr dataset and then ask users to rate, in a four-point scale, whether the recommendations are appropriate to a given image. Similarly, De Meo et al. [4] extend the annotations of Delicious' bookmarks[2] and then ask users to evaluate the relevance of every tag/resource association. Jäschke et al. [13] perform a small evaluation based on a real-world scenario where users have to tag bookmarks in BibSonomy[3]. Specifically, precision and recall metrics are computed by comparing tag recommendations performed to every bookmark and the final taglines that users introduced. Due to its subjectiveness and many different ways to be accomplished, tag recommendation is not an easy task to evaluate, and some advantages and disadvantages can be found in both user-based and information retrieval evaluation approaches [8]. However, there is a clear lack of user-based evaluation in previous work, and we believe that every recommendation system should be validated at some point using both evaluation strategies. Proper user feedback should be helpful not only to compare tag recommendation methods but also to better understand the nature of the task and learn how can systems be improved.

The contribution of the present work is twofold. First, we propose an extended version of the best performing tag recommendation method found in our previous work [7]. The main idea behind this extended method is to exploit the automatic classification of the resources to be annotated into a number of predefined classes to further adapt the tag suggestions to the context of these classes. This classification is based on the tags that users start introducing during the annotation process. In this way, instead of personalising recommendations for particular users, we "personalise" them to particular classes of resources. Next, as a second contribution, we perform a comprehensive user-based evaluation through an online experiment were participants are presented with some resources which have to be annotated with the help of a tag recom-

[2]`www.delicious.com`
[3]`www.bibsonomy.org`

mendation system. These kind of user-based evaluations are very costly and we have seen that they are not very common in the tag recommendation literature. For that reason, we believe our contribution is of great valuable to the community. In our evaluation, we compare the recommendation method we proposed in previous work and the extended version we describe here along with two random baselines. Moreover, we perform a complementary evaluation based on a tag-prediction task following common information retrieval methodologies. In our previous work [7], the tag recommendation methods were evaluated using a tag-prediction task and compared favourably against four baselines and two state of the art methods [8, 22]. For this comparison, we used data from the folksonomies of Freesound[4], an online audio clip sharing site with more than 3,5 million registered users and 180,000 uploaded sounds [5], and Flickr. Therefore, the recommendation methods were tested in the audio and image domains. Similar results were obtained in both scenarios. In this work, evaluations are carried out in the context of Freesound. Results show that the newly proposed recommendation method brings a statistically significant improvement over the previous method, according to both user-based and prediction-based evaluations. Analysing user-based evaluation results we find that participants which are experienced in working with sound libraries tend to better appreciate the improvements of the new tag recommendation method we describe here. Moreover, we see that the more familiarised the users are with Freesound, the more the number of tag suggestions they accept as valid annotations. User feedback reveals that tag recommendation methods tend to be more useful when recommending broad tags (i.e., referring to generic concepts). Participants also recognise tag annotation as a particularly difficult task, specially if the resources being annotated are not authored by themselves.

The rest of the paper is organised as follows. First, we summarise the steps of the tag recommendation method we proposed in previous work and describe the new approach based on the classification of input tags (Sec. 2). Then, we describe the online experiment we designed for user-based evaluation (Sec. 3). Results of the online experiment are reported in Sec. 4, and the complementary prediction-based evaluation is described and reported in Sec. 5. We conclude the paper with a discussion about our findings and future work (Sec. 6).

2. Tag recommendation methods

The two tag recommendation methods we describe in this work are based on tag-tag similarities derived from the folksonomy of Freesound. Given a set of input tags Γ_{I}, the methods output a set of recommended tags Γ_{R}.

2.1. General tag recommendation

The general tag recommendation method presented in [7], which we denote by Gen, consists of three steps (Fig. 1):

[4] `www.freesound.org`

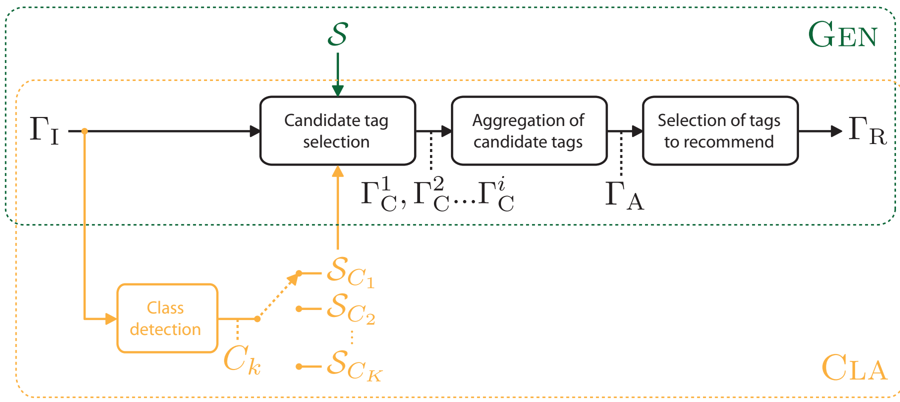


Figure 1: Schematic block diagram of the general (Gen) and class-based (Cla) tag recommendation methods.

1. Candidate tag selection: Given a set of input tags Γ_I, this step uses a tag-tag similarity matrix $\mathcal{S}$ derived from the Freesound folksonomy to select a set of N candidate tags Γ_C^i for each input tag Γ_{I_i}. The tag-tag similarity matrix $\mathcal{S}$ is constructed by computing the association matrix $\mathcal{D} = \{d_{i,j}\}$, which represents the associations between tags and audio clips in the Freesound folksonomy ($d_{i,j} = 1$ if audio clip a_i is labeled with tag t_j, and $d_{i,j} = 0$ otherwise). Hence, $\mathcal{D}$ is a sparse matrix that has as many columns as audio clips in Freesound and as many rows as the set of distinct tags being used to label these audio clips[5]. Given $\mathcal{D}$, the tag-tag similarity matrix is obtained as $\mathcal{S} = \mathcal{D}\mathcal{D}'$ ($'$ indicates transposition), and we apply a simple normalisation to the elements $\left\{s_{t_i,t_j}\right\}$ of $\mathcal{S}$ so that s_{t_i,t_j} corresponds to the cosine similarity between tags t_i and t_j on the basis of their co-occurrence in audio clips. Tags in Γ_C^i are selected as the N most similar tags to a given input tag Γ_{I_i}.
2. Aggregation of candidate tags: Given the sets Γ_C^i from the first step, candidates are assigned a score ϕ and aggregated into a single list of tags with scores Γ_A. Such score is determined by the candidate similarity-based ranking so that $\phi = 1$ for the most dissimilar candidate to a given input tag and $\phi = N$ for the most similar one. The scores of tags that are present in different sets of candidates Γ_C^i are added when aggregated in the final set Γ_A.
3. Selection of tags to recommend: Considering the scores in Γ_A, this step

[5]In order to reduce the computational cost of the operations performed in this step and to get rid of potentially noisy tags, when building the association matrix we only consider tags whose frequency of occurrence is higher that a threshold $\omega = 10$ (i.e. we only consider tags that are used at least 10 times in the Freesound folksonomy). In this way the number of rows of the association matrix is reduced by ≈80%, with only around ≈10% of the associations between tags and audio clips being actually ignored [7].

determines a threshold ϵ to select the tags that are finally recommended. Here we use the strategy of determining the threshold ϵ as a percentage of the maximum score in Γ_{A} [7]. Tags in Γ_{A} are sorted by their score and those that satisfy $\phi >= \epsilon$ are outputted as Γ_{R}, the final set of recommended tags.

In this way, the method Gen can generate a sorted list of recommended tags Γ_{R} given a set of input tags Γ_{I} and a tag-tag similarity matrix $\mathcal{S}$ which is derived from previous tag associations. Given that this method does not take into account any audio-specific information such as content features, it is general enough to be applied to other kinds of multimedia domains. Example applications for audio and images, as well as more detailed explanations, are provided in [7].

2.2. Class-based tag recommendation

The proposed class-based tag recommendation method, which we refer to as Cla, is a variation of Gen based on the classification of the input tags Γ_{I} into a set of K predefined audio classes. For every class C_k, $k \in [1, K]$, a tag-tag similarity matrix $\mathcal{S}_{C_k}$ is built in the same way as in the Gen method, except that in this case only the tag assignment information corresponding to the audio clips of the current class is considered (see below). As a result, a different tag-tag similarity matrix can be computed for every audio class, and the matrix $\mathcal{S}_{C_k}$ that is used in the candidate tag selection step of the recommendation process depends on the classification of the input tags Γ_{I} (Fig. 1). Once the candidates are selected, the other two steps (aggregation of candidate tags and selection of tags to recommend) are computed exactly in the same way as in Gen.

2.2.1. Classification of input tags

The classification of input tags is performed using a supervised learning model trained with the original tag annotations of audio clips in Freesound. We defined $K = 5$ audio classes (Table 1) and manually built a ground truth of 1,200 audio clip examples of each class. Then, we trained a multivariate Bernoulli Naive Bayes classifier feeding it with the taglines of the audio clips in the ground truth. Given a set of input tags Γ_{I}, the classifier can predict which category C_k better fits the input. Details on the class detection step and the process we followed for defining the audio classes, building the ground truth and evaluating the classifier can be found in [6]. The resulting classification system is able to classify a set of input tags Γ_{I} within the five defined classes with different accuracies depending on the length of Γ_{I}. The lowest accuracy, obtained when $|\Gamma_{\mathrm{I}}| = 1$ (i.e., only one tag is given to the classifier), is approximately 75%. For $|\Gamma_{\mathrm{I}}| \geq 4$ the classification accuracy reaches a plateau between 90 and 95%.

2.2.2. Computation of tag-tag similarity matrices

As mentioned, the process of building the tag-tag similarity matrices $\mathcal{S}_{C_k}$ is the same as the one for building $\mathcal{S}$, except that for every matrix $\mathcal{S}_{C_k}$ we only consider tag assignment information from audio clips belonging to C_k. For that

Class name	Description and examples
SoundFX	Sound effects (including *foley*), footsteps, opening and closing doors, alarm sounds, cars passing by, animals, and all kinds of noises or artificially created glitches.
Soundscape	Environmental recordings, street ambiances or artificially constructed complex soundscapes.
Sample	Instrument samples including single notes, chords and percussive hits (e.g. single notes of a piano recorded one by one and uploaded as different audio clips, or samples from a complete drum set).
Music	Musical fragments such as melodies, chord progressions, and drum loops. This class is to Sample what Soundscape is to SoundFX.
Speech	All sorts of speech-related audio clips such as text reading, single words or recordings of text-to-speech processors.

Table 1: Audio classes.

we reused the classification system described in Sec. 2.2.1 to classify all audio clips in Freesound in one of the five audio classes, with input tags corresponding to the original taglines of audio clips in Freesound. Then, matrices $\mathcal{S}_{C_k}$ can be built by only considering the columns of $\mathcal{D}$ corresponding to the audio clips of C_k. Hence, $\mathcal{S}_{C_k} = \mathcal{D}_{C_k}\mathcal{D}'_{C_k}$, where $\mathcal{D}_{C_k}$ is a subset of $\mathcal{D}$ where the columns corresponding to audio clips not in C_k are removed. Each matrix $\mathcal{S}_{C_k}$ is normalised using the same process we use for $\mathcal{S}$ (Sec. 2.1).

Notice that the similarity value between two tags t_i and t_j will be different in every matrix $\mathcal{S}_{C_k}$ and in $\mathcal{S}$, with $\mathcal{S}_{C_k}$ being tailored to the particular context of the k-th class. Also notice that the number of distinct tags resulting from considering all audio clips belonging to C_k will be smaller than the total number of distinct tags resulting from considering all audio clips from all classes (the size of the *class vocabulary* will be smaller than the size of the *general vocabulary*). Therefore, there will be some "all-zeros" rows in $\mathcal{S}_{C_k}$, corresponding to the tags that are not used in the context of the particular class C_k. Hence, these tags are never recommended when using $\mathcal{S}_{C_k}$.

3. User-based evaluation

We designed an online experiment where participants have to tag a set of audio clips from Freesound with the help of the tag recommendation systems of Sec. 2. The experiment was online for 15 days during June 2013, and was publicised in the Freesound front page. The goal of this experiment is twofold. First, we want to assess which of the recommendation methods is more useful for users when tagging audio clips. Second, we want to get qualitative user feedback to better understand the strengths and weaknesses of the considered tag recommendation systems and, in a further stage, to understand the potential strengths and weaknesses of tag recommendation processes in general. As

Freesound dataset	
Number of audio clips	140,622
Number of unique tags†	43,696
Number of contributor users‡	6,948
Number of tag assignments	990,574
Average tags per audio clip (tagline length)	7.044

Tag-tag similarity matrices	Num. audio clips	Vocabulary size
General matrix ($\mathcal{S}$)	140,622	7,710
Matrix for class SoundFX	29,725	4,584
Matrix for class Soundscape	38,001	5,768
Matrix for class Sample	26,452	3,280
Matrix for class Music	34,139	4,303
Matrix for class Speech	15,305	3,557

Table 2: General statistics of the Freesound dataset and the resulting tag-tag similarity matrices. †Some of these tags are not semantically unique, and may include synonyms and typographic errors. ‡Users that have contributed by uploading at least one audio clip.

mentioned in Sec. 1, this is yet an under-explored area.

Along with Gen and Cla, we also evaluate two random variants of them, named RGen and RCla, respectively. These differ from the original variants in that, in the final step of the recommendation process, the set of recommended tags Γ_R is replaced with an alternative set of the same length containing randomly selected tags either from the general vocabulary (RGen) or from the corresponding particular class vocabulary (RCla). Notice that the general vocabulary is always bigger than any of the individual classes' vocabulary. Hence, the random selection in RGen is performed over a bigger and more diverse pool of tags. Participants were not aware of the particular recommendation method underlying tag suggestions nor knew about the five audio classes in which we classify all annotated audio clips. The dataset we use for the evaluation comprises Freesound data[6] gathered between April 2005 and May 2012 (Table 2). It includes tag assignment information which relates tags, audio clips and users, and it is used to build the tag-tag similarity matrices $\mathcal{S}$ and $\mathcal{S}_{C_k}$, as explained in Sec. 2.2.2.

The online-experiment proceeded as follows:

Instructions page: First, participants were presented with an introduction page displaying detailed instructions for the experiment (Fig. 2). Participants were told they would have to annotate 20 audio clips from Freesound, using as many tags as they felt appropriate for every clip (we suggested participants to use five or more tags, but it was not mandatory). Participants were also told that as soon as they started typing tags, a list of

[6]Freesound data, including audio clips and tag annotations, can be gathered using the pubic Freesound API (www.freesound.org/help/developers/).

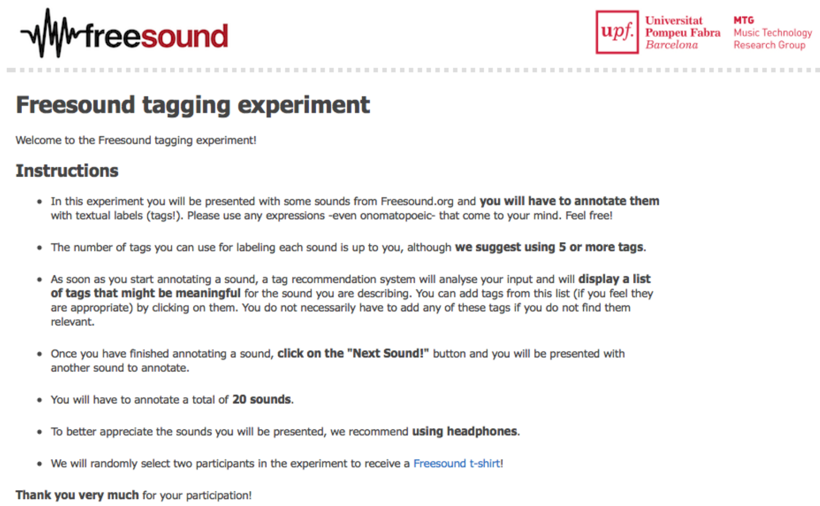


Figure 2: Screenshot of the instructions page.

tag suggestions would appear and that they could choose tags from this list if they felt the suggestions were appropriate. We also recommended participants to use headphones for better listening conditions.

Questionnaire: After the introduction, a short questionnaire (Fig. 3) was presented to collect some basic user data and information about their experience in working with sound libraries, their experience using Freesound (including the number of uploaded sounds) and their native language (in particular to be able to differentiate between native and non-native English speakers).

Audio clip annotation: Once the questionnaire was completed, participants started annotating audio clips. From the ground truth we defined when designing the recommendation system (Sec. 2.2.1), we manually selected 50 audio clips per class[7]. These clips were selected trying to cover a certain variety of sounds and avoiding those that would presumably be very hard to annotate. From this pool of 250 clips, every participant was assigned a random selection of four clips per class. Then, each of the four clips was assigned a different tag recommendation method that would be used when the participant annotated the clip. In this way, every participant was assigned a total 20 audio clips, equally distributed among audio classes and recommendation methods. Participants were presented with the first audio clip and had to annotate by typing tags in a text box.

[7]The clips we selected for the annotation phase of the online experiment (a total of 250, 50 per class) were removed from the ground truth and thus were not used to train the classifier described in section 2.2.1.

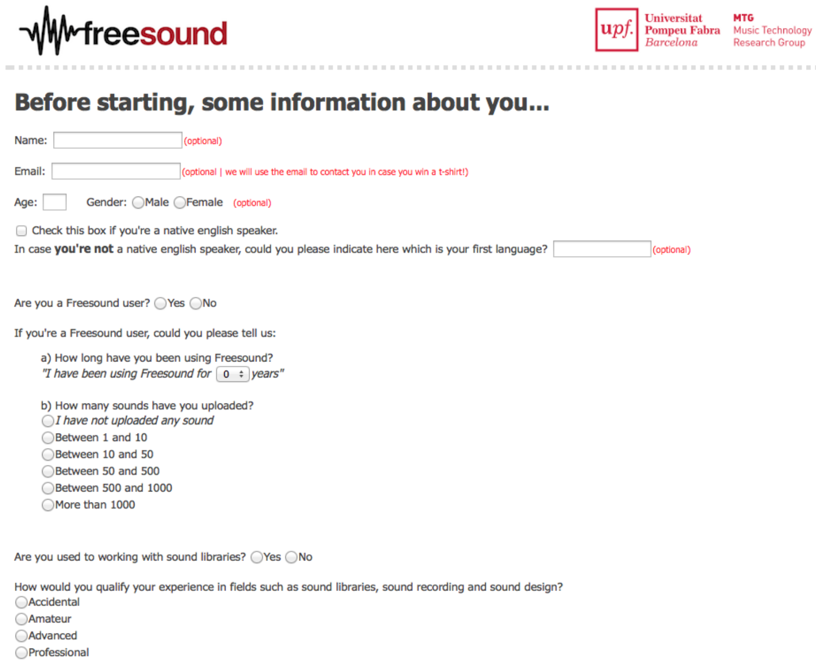


Figure 3: Screenshot of the questionnaire page.

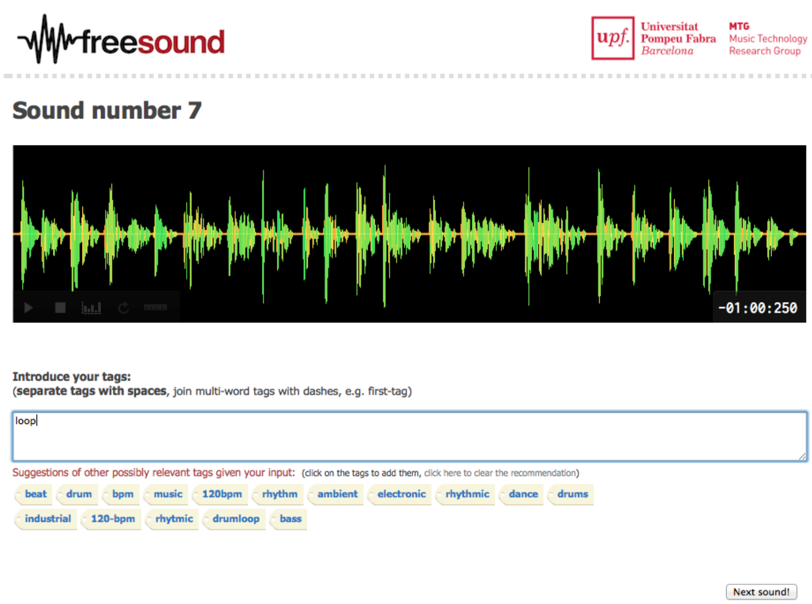


Figure 4: Screenshot of the sound annotation page.

The audio clip could be reproduced using a web player that also showed a visualisation of the waveform and the spectrogram of the audio clip. As soon as the participant started typing, a list of suggested tags appeared below the text box. This list was computed using the tag recommendation method assigned to the currently annotated audio clip, and was being updated every time a new tag was written in the text box[8]. Users could click over the tags shown in the list to automatically append them in the text box (Fig. 4). Once a participant considered an audio clip was fully annotated, she could click on the "Next sound" button and be presented with the following clip. Participants were also provided an URL that they could save for later resuming the experiment in case they did not want to annotate all clips in one go. Noticeably, we logged information about all the keystrokes and mouse clicks that participants performed with the corresponding timestamps.

Feedback page: After annotating the 20 audio clips, participants were presented with a page thanking their participation and offering some space in a text box to give some feedback about the experiment. Alternatively, they were also offered to write the feedback in a particular section of the Freesound forums.

Considering the logs resulting of the user experiments we define a simple measure for evaluating the "usefulness" of every tag recommendation method in the tagging process. The measure consists in counting, for every set of tags assigned to an audio clip by a particular participant, the number of these tags that were recommended by the system during the annotation process (i.e., the number of recommended tags that were *accepted* by the participant). Let Γ_{P} be the set of tags that a participant used to annotate a particular audio clip, and let Γ_{R_m} be one of the sets of recommended tags that were presented to the user in the successive M tag recommendations during the tagging process of that particular audio clip. Then, we can define Λ, the number of accepted tags, as

$$\Lambda = \left| \Gamma_{\mathrm{P}} \cap \left(\bigcup_{m=0}^{M} \Gamma_{\mathrm{R}_m} \right) \right|, \tag{1}$$

where $|\ \ |$ measures set cardinality. Notice that Λ is roughly equivalent to a standard recall measure (without the normalization by $|\Gamma_{\mathrm{P}}|$). We employ this measure instead of standard precision and recall (e.g., as done in [13]) because the nature of our evaluation has some particularities which make such metrics less useful. As described above, in our evaluation system several tag recommendations are performed during the annotation of a single clip (i.e., every time that a new tag is introduced the recommendation is recomputed). As a

[8]Similarly to the Freesound upload system, tags had to be separated by spaces and multiwords joined with dashes. Hence, the recommendation was updated every time a blank space was introduced.

result, the total number of recommended tags for every audio clip is much larger than the final number of assigned tags. If we computed precision and recall by comparing the whole set of recommended tags for every audio clip with the final taglines assigned by users, we would obtain very low precision values which, in our opinion, are not as representative as Λ. In our evaluation (and in a real-world tag recommendation scenario), users are the ones who finally decide which of the recommended tags are relevant for a particular resource. Therefore, the length of the recommendation is not as important as the fact that it contains meaningful suggestions (i.e., recall is much more important than precision).

4. Results

During the two weeks the experiment was online we gathered a total 201 experiment logs from 190 unique participants (some participants decided to repeat the experiment more than once). Among all these experiment logs, 80 correspond to unfinished experiments (i.e., with less than 20 audio clips annotated) which we do not consider in the analysis. In addition, we apply a filter to discard logs from experiments that were finished very quickly and with very few calls to the recommendation methods. More specifically, we discard logs from experiments completed in less than 10 minutes (average of 30 seconds per audio clip) and from experiments not reporting a minimum of three calls to the recommendation system for every annotated audio clip. We discard these logs as we consider that participants did not pay enough attention when annotating audio clips and thus contain potentially noisy data. After filtering, we are left with 70 logs that we consider as sufficiently reliable data for analysis. In the following subsections we show the results of different aspects of the online experiment analysis.

4.1. Accepted tags per recommendation method

First, we report on the basic accuracy of the considered tag recommendation methods (Table 3, leftmost column). We observe that random methods RCla and RGen report way lower average Λ than Cla and Gen. Thus, our methods do perform much more meaningful recommendations than the random baselines. Interestingly, we also observe that both class-based methods Cla and RCla report higher averages than their general counterparts Gen and RGen. This suggests that tag recommendations improve when using class-based methods. However, the differences are not statistically significant[9].

Next, we repeat the same analysis but considering different groups of experiment logs according to the questionnaire that participants had to fill at the

[9] If not stated otherwise, statistical significance is assessed by performing pairwise comparisons using the Mann-Whitney U test with α=0.05 [17]. When performing multiple comparisons, we apply a correction to the rejection criteria in order to reduce the familywise error rate. In particular, we use the Holm-Bonferroni correction [11]. Notice that these are robust and stringent criteria for measuring statistical significance (cf. [21]).

	All	Expert	Non-expert	Native	Non-native
Cla	2.414 (2.775)	2.547 (2.988)	2.179 (2.224)	2.950 (3.382)	1.963 (2.027)
Gen	2.154 (2.526)	2.163 (2.663)	2.147 (2.229)	2.656 (3.006)	1.732 (1.938)
RCla	0.260 (0.671)	0.278 (0.680)	0.211 (0.663)	0.300 (0.705)	0.226 (0.638)
RGen	0.166 (0.455)	0.139 (0.458)	0.253 (0.458)	0.194 (0.518)	0.142 (0.392)

Table 3: Average number of accepted tags Λ (standard deviation into parenthesis) of the user-based evaluation approach for the following groups of participants. From left to right these correspond to all participants, expert participants, non expert participants, native English speakers and non-native English speakers.

beginning of the experiment (Table 3). In particular, we compute Λ for each recommendation method considering groups of logs corresponding to experienced participants (i.e., participants that checked the box marked with the question "Are you used to working with sound libraries?" in the questionnaire; second column in Table 3), non-experienced participants (third column), native English speakers (fourth column), and non-native speakers (fifth column). We again observe that, except for RCla and RGen in the non-expert group, all class-based methods report higher averages than the general methods. This further supports the idea that class-based recommendations bring some improvements over the general method. Interestingly, in the case of experienced participants, the difference between Cla and Gen increases with respect to the same comparison when considering all participants. In this case we get a statistically significant increase of 0.38 ($p < 2.91 \cdot 10^{-2}$). Furthermore, the difference between RCla and RGen also increases for the experts (with respect to all participants) and becomes statistically significant ($p < 2.47 \cdot 10^{-3}$). This suggests that expert participants clearly appreciate a difference between Cla and Gen methods (even for the random versions) and find class-based recommenders to be more useful. On the other side, we observe that when analysing the non-experienced participants group, the differences between class-based and general methods gets blurred, with almost no difference between the two types of recommendation methods. Thus, non-experienced participants are not able to tell the difference between class-based and general recommendations. Overall, these results indicate that the usefulness of class-based tag recommendations compared to general recommendations is slightly higher, and specially in the case of experienced participants.

Considering the last two groups of participants (native and non-native English speakers), we observe that the differences between class-based and general recommendation systems are quite similar to those obtained when considering all participants. Class-based systems report a higher Λ but the increments are practically the same for both native and non-native groups (there is no statistically significant difference between the increments). Thus, we do not see a direct general implication of language in method preference. Nevertheless, there is a significant difference in the absolute number of accepted tags among the native and non-native participant groups (Table 3). Native English speakers tend to accept an average of 0.96 tags more than non-native ones ($p = 4.61 \cdot 10^{-3}$). Fur-

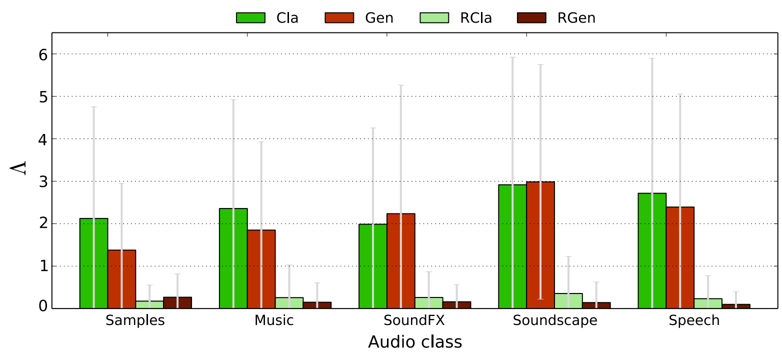


Figure 5: Average accepted tags (Λ) per audio class and recommendation method.

thermore, we observe that native English participants tend to annotate audio clips with an average of 0.32 tags more than non-native ones ($p = 3.24 \cdot 10^{-6}$). Thus, in our experiments, native speakers consistently use more tags for describing audio clips than non-native speakers and tend to accept more recommendations. To the best of our knowledge, this is the first time evidence is reported with regard to the comparison of native's and non-native's tagging behaviour. Our results suggest that native speakers use more tags when describing online resources than non-native participants and that, therefore, this aspect should not be overlooked in future studies. Overall, we see that both native and non-native speakers prefer Cla over Gen (and RCla over RGen), but that this preference is not stronger than in any of the other user groups.

4.2. Accepted tags per audio class

To gain insight about how do recommendation methods work for the different audio classes defined above (Table 1), we grouped annotated sounds by class and recommendation method and computed the average number of accepted tags Λ for each group (Fig. 5). In general, clips under Soundscape and Speech classes reported higher Λ than clips under the other classes. This is probably because there are some tags such as `field-recording`, `nature` or `voice` which are very common in these classes and are very generic (i.e., could be used to annotate almost any clip in Soundscape or Speech classes).

It can be also observed that not all audio classes feature a higher Λ for the Cla method than for the Gen method. Soundscape clips report higher Λ for Gen than for Cla, although the difference of 0.07 is not statistically significant ($p = 4.56 \cdot 10^{-1}$). SoundFX clips also report higher Λ for the Gen method and, although the difference is still not statistically significant ($p = 3.80 \cdot 10^{-1}$), the increase of 0.25 is this time bigger. Sample, Music and Speech classes report higher Λ for Cla recommendations, with larger Λ increases and with improved statistical significance. This suggests that the knowledge-based adaptation that the Cla performs is better exploited in Speech, Music and Sample classes than in Soundscape or SoundFX. We hypothesise that the vocabulary needed to accurately describe clips from the former classes is more reduced than the vocabulary needed for other audio clips. Therefore, the class-based method can easily adapt to the class context and produce better recommendations, probably including less generic tags than the ones that would be recommended using the general method. On the other side, clips under SOUNDSCAPE and SOUNDFX classes cover a wider range of sounds and need a larger vocabulary to be well-described. In this situation, the CLA method does not adapt well and does not improve the GEN results. Our hypothesis is partially supported by looking at the actual size of the resulting class vocabularies after computing the tag-tag similarity matrix per class ($\mathcal{S}_{C_k}$, Table 2). SPEECH, MUSIC and SAMPLE produce smaller similarity matrices, with less tags in the vocabulary, than SOUNDSCAPE and SOUNDFX.

4.3. Correlation between number of uploaded sounds and accepted tags

All participants in our experiment were Freesound users. However, not all of them had experience in uploading and tagging audio clips in Freesound. In order to get some insight as how being used to tagging audio clips affects Λ, we computed the correlation[10] between the number of uploaded sounds and the number of accepted tags, grouping audio clips into the four evaluated recommendation methods (Table 4). We find the strongest correlation for the CLA method ($\varrho = 0.276$, $p < 3.76 \cdot 10^{-7}$). Thus, in this case, Λ tends to grow along with the number of uploaded sounds. A less significant correlation is reported for the GEN method ($\varrho = 0.105$, $p < 5.61 \cdot 10^{-3}$). RCLA and RGEN present no significant correlations ($\varrho = 0.087$, $p < 1.13 \cdot 10^{-1}$ and $\varrho = 0.063$, $p < 2.55 \cdot 10^{-1}$, respectively). This finding suggests that the more familiar the participants are with the Freesound uploading and tagging process, the more recommended tags they tend to accept, specially when recommendations are generated with the CLA method. This result is consistent with the previous observation that experienced participants tend to accept more tags than non-experienced ones when recommendations are generated by CLA (Sec. 4.1). Again, we are not aware of any study considering user familiarity in the context of resource tagging. Therefore, our results represent a novel and original contribution with regard to this aspect.

4.4. Timing aspects

Timing is also an often unconsidered aspect when evaluating tag recommendation systems. However, it is interesting because it can reveal some insights about the annotation process. We measured the average time invested for annotating an audio clip in our experiments and observed that there exists a significant correlation between the length of the audio clips and the time invested to annotate them, being shorter clips the fastest to describe ($\varrho = 0.24$,

[10] We employ the Spearman's rank correlation coefficient [10], with ϱ denoting the correlation coefficient and p the p-value associated with it.

Number of uploaded sounds†	COM	GEN	RCOM	RGEM
0	2.105	2.036	0.221	0.126
1 to 10	1.823	2.027	0.293	0.133
11 to 50	2.580	1.820	0.220	0.240
51 to 500	2.289	2.222	0.311	0.133
501 to 1000	4.160	2.035	0.380	0.300

Table 4: Average number of accepted tags Λ per number of uploaded sounds and recommendation method. †The ranges in the number of uploaded sounds are determined in the questionnaire that participants had to fill at the beginning of the experiment (Fig. 3).

$p < 5.68 \cdot 10^{-19}$). That could be expected, as shorter clips tend to be less complex and need less time for listening to them. Consistently, audio clips belonging to the SOUNDSCAPE class need an average of 15 extra seconds to be described when compared to clips belonging to other classes ($p < 8.12 \cdot 10^{-3}$). On the other side, SAMPLE clips need less time than the rest ($p < 3.15 \cdot 10^{-2}$). This can be explained because SOUNDSCAPE clips are generally longer than clips from other classes, while SAMPLE clips tend to be shorter. We have not observed any statistically significant differences in the average time invested in annotating audio clips when comparing the four different recommendation methods. Therefore, the choice of a recommendation method does not seem to affect the time needed to annotate audio clips.

4.5. User feedback

In the last phase of the online experiment, participants were provided the opportunity to give some feedback in the form of comments (Sec. 3). We observe some recurring opinions that, if extrapolated, bring also valuable insights into recommendation processes in general. First of all, participants agree in that the process of annotating audio clips (and by extension the process of recommending tags) is a very hard task, and that recommendations are a generally useful tool but not always needed or used. In our case, the 30% of all tag annotations performed during the experiment were suggested by the recommendation systems[11].

A lot of participants point out that annotation is especially hard when the audio clip being described is not recorded/created by the person annotating it (which was always the case in our experiment). In those cases, there is a lot of meaningful information about the sound which most of the times can not be determined without the knowledge of how the clip was created (e.g., software used, recording device, location of a recording, etc.). Some participants also point out that in order to perfectly annotate musical audio clips such as drum loops or instrument notes, a lot of time needs to be invested in determining properties such as beats per minute or the pitch of a note. Those issues are particularly relevant in our context, where participants had to annotate audio clips

[11]This percentage is computed without taking into account tag recommendations performed with random methods which obviously did not provide meaningful recommendations.

not created by themselves. Finally, another repeated comment is that tag suggestions are more useful for “nature” and “human-related” audio clips, whereas “abstract” and “synthetic” clips require more tags to be manually introduced before some meaningful suggestions are made. These comments are somehow aligned with the results reported in Fig. 5, where we see that SOUNDSCAPE and SPEECH classes are the ones that report higher Λ.

4.6. Tag analysis

We perform here a close-look analysis to the experiment logs in order to get some insight in the type of tags that are recommended and in which cases those are accepted by participants. We detect several interesting patterns that we believe also help comprehending in more detail tag recommendation processes in general. First of all, there are some tags which are recommended and accepted a lot of times in the online experiment. These tags correspond to very generic concepts such as `field-recording`, `voice`, `electronic`, `loop`, `nature` or `percussion`. These recommendations are useful to provide some kind of general categorization to annotated audio clips, but clips only tagged with these kind of tags do clearly lack specificity in the annotations. We observe that another recommendation pattern consists in tags that are suggested many times but are rarely accepted. This is the case of tags such as `sound` or `recording`, for which we hypothesise that the meaning is too obvious to be considered as relevant information for participants. It is also the case of tags like `soundscape`, `percussion-loop`, `drum-loop` or `natural-reverb` which are normally represented by alternative tags (or combinations of tags) such as `field-recording` (instead of `soundscape`), `loop`, `percussion`, `drum`, `natural` or `reverb`.

We also observe that there are some tags with low acceptance because of its subjective meaning (e.g. `groovy`, `threatening`) or because participants can not assess its correctness because they are not the authors of the annotated clips (e.g. `multi-sample`, `improvised`). Obviously there are also some suggested tags which are not accepted because they are simply not appropriate for the clips being described. That could be the case of tags like `piano`, `guitar` or `pad` which are sometimes recommended to audio clips which clearly do not contain piano, guitar or pad-like sounds. Finally, we observe a last group of suggestions which correspond to tags not usually suggested but normally accepted such as `annoucement`, `synthesizer`, `footsteps` or `airplane`. We consider these as being very good recommendations as they correspond to not-so-general concepts and are apparently recommended only when they are needed. Overall, recommendations provided by our methods tend to be useful when recommending general tags, referring to concepts than can be used as a broad categorisations of the audio clips. However, recommendations are not as useful when they refer to more detailed aspects of the sounds being annotated.

5. Complementary evaluation

In order to complement the performed user-based evaluation, we also consider a more systematic and empirical assessment of the different tag recommendation methods (Cla, Gen, RCla and RGen) following the methodology we described in [7]. This complementary assessment follows a typical information retrieval evaluation setup based on a tag prediction task which we now describe.

5.1. Prediction-based evaluation methodology

For this evaluation we consider audio clips and annotations of the same Freesound dataset described in Sec. 3. We perform a 10-fold cross-validation following the methodology described by Salzberg [21] and others. For each fold, there is a training phase consisting of two steps which preprocesses all the necessary data for performing recommendations with the four evaluated methods. The first step consists in training a classifier that allows the classification of the input tags in one of the five defined audio classes, as described in Sec. 2.2.1. To do that, we feed the classifier only with these audio clips that are present both in the training set and in the ground truth we built when designing the system (i.e., we only use audio clips from the training set that we know to which audio category they belong to).

The second step of the training phase consists in building the general tag-tag similarity matrix $\mathcal{S}$ and the matrices $\mathcal{S}_{C_k}$ for every class C_k. For that we use information from all the audio clips in the training set. Notice that building $\mathcal{S}_{C_k}$ requires the classification of all audio clips of the training set in one of the five defined categories (Sec. 2.2.2). We perform that classification using the same classifier trained in the first step of the training phase. Hence, this classifier is not only used in the recommendation process to classify the input tags and select a similarity matrix $\mathcal{S}_{C_k}$, but it is also used to build the similarity-matrices $\mathcal{S}_{C_k}$ by classifying the audio clips of the training set.

After the training phase, we pick every audio clip in the evaluation set and randomly delete a set of tags Γ_{D} from its originally assigned tags, yielding Γ_{I}, the input to our recommendation system. The number of tags we delete is chosen uniformly at random, with the only constraint of leaving a minimum number of input tags of $|\Gamma_{\mathrm{I}}| \geq 3$ so that there is presumably enough information for the recommender systems to provide good recommendations [7]. This constraint also implies that in order to be able to remove at least one tag for each audio clip ($|\Gamma_{\mathrm{D}}| \geq 1$), we can only consider for evaluation the audio clips that have at least four tags[12]. After we remove some tags, we run the four tag recommendation methods using Γ_{I} as input and the similarity matrices we computed in the training phase.

As evaluation measures we compute standard precision (P_n), recall (R_n), and f-measure (F_n) for each individual audio clip n according to

$$P_n = \frac{|\Gamma_{\mathrm{R}} \cap \Gamma_{\mathrm{D}}|}{|\Gamma_{\mathrm{R}}|}\ ,\ R_n = \frac{|\Gamma_{\mathrm{R}} \cap \Gamma_{\mathrm{D}}|}{|\Gamma_{\mathrm{D}}|}\ ,\ \text{and } F_n = \frac{2P_nR_n}{P_n + R_n},$$

[12]This filtering is done before the whole evaluation process starts, therefore we evaluate the same number of clips in each fold.

where Γ_{R} is the set of recommended tags and Γ_{D} is the set of deleted tags. Then, global P, R and F measures for each tag recommendation method are calculated by averaging P_n, R_n and F_n across all resources $n \in [1, N]$ evaluated with the chosen recommendation method.

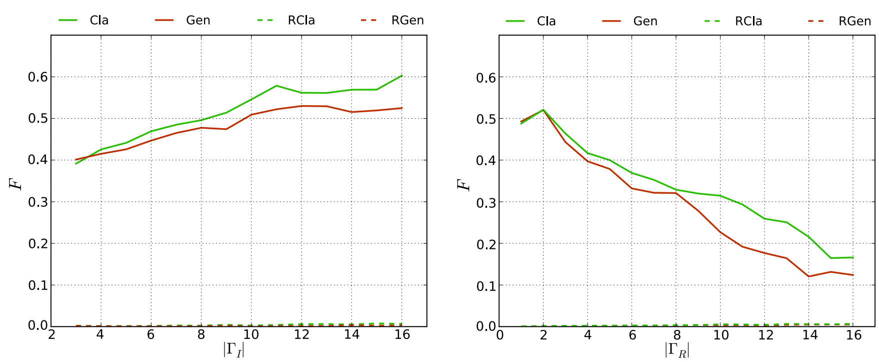
The prediction-based evaluation approach is interesting as it allows us to evaluate the different recommendation methods in a systematic way and using a lot of audio clips. In previous work [7], we used this evaluation methodology to exhaustively compare eight variations of the Gen recommendation method (using different sets of parameters for each one of the recommendation steps) against four baselines and two state of the art folksonomy-based tag recommendation methods, and using data from the folksonomies of Freesound and Flickr. That number of methods could have not been comprehensively compared through a user-based evaluation approach such as the one presented in the above sections. However, prediction-based evaluation has an important limitation which is that we need an extensive ground truth to evaluate whether our predictions are correct or not. In our case, this ground truth is composed by the original taglines of sounds in Freesound. This means that the recommendations we evaluate will only be considered as "correct" recommendations if they contain tags that the original author of the sound used to annotate it. As a result, tags that could be subjectively considered as good recommendations for a particular audio clip but are not present in the original annotations do not count as correct predictions. Moreover, prediction-based evaluation does not allow the collection of qualitative user feedback that, as we have seen before, can shed some light on relevant aspects of the recommendation process. For that reason, we state that the prediction-based evaluation approach may be taken as a complement to the results already described in the previous sections, allowing us to further test our previous findings.

5.2. Prediction-based evaluation results

Results for the four evaluated tag recommendation methods appear to be very similar to what we observe in the user study (Table 5). We can see that Cla outperforms Gen by a small but statistically significant difference of 0.011 ($p < 6.51 \cdot 10^{-8}$). This difference suggests that Cla can successfully take advantage of the classification step and the knowledge derived from the ground truth to slightly improve the recommendations of the system. As expected, random methods RCla and RGen score much lower F than Cla and Gen. Nevertheless, it is interesting to note that RCla also features a statistically significant increase in F with respect to RGen ($p < 1.57 \cdot 10^{-24}$). This increase can be explained by recalling that the pool of tags from which the random selection is performed in RCla is different in every audio class and it always contains less tags than the pool in RGen (Sec. 2.2.2). Hence, these results suggest that at least some tags which are not relevant for a particular audio class are effectively removed when building the similarity matrices $\mathcal{S}_{C_k}$. We also observe that Cla and Gen feature a very similar number of recommended tags $|\Gamma_{\mathrm{R}}|$, with an average of 3.99 and 3.88 tags, respectively.

	P	R	F
Cla	0.476 (0.428)	0.488 (0.424)	0.440 (0.389)
Gen	0.486 (0.429)	0.467 (0.408)	0.429 (0.372)
RCla	0.003 (0.031)	0.003 (0.038)	0.002 (0.025)
RGen	0.002 (0.024)	0.002 (0.031)	0.001 (0.019)

Table 5: Average precision, recall and f-measure (standard deviation in parenthesis) for the prediction-based evaluation approach. Results are sorted by f-measure.



(a) (b)

Figure 6: Average f-measure F as a function of the number of input tags $|\Gamma_\mathrm{I}|$ (a) and the number of recommended tags $|\Gamma_\mathrm{R}|$ (b).

If we analyse F as a function of the number of input tags $|\Gamma_\mathrm{I}|$ and the number of recommended tags $|\Gamma_\mathrm{R}|$ we can get some more insight on the behaviour of the considered recommendation methods (Fig. 6). For instance, we see that both Cla and Gen have a tendency of increasing F as the number of input tags also increases (Fig. 6(a)). This suggests that the recommendation system is able to provide better recommendations when it is feed with more input tags. The opposite happens with the number of recommended tags (Fig. 6(b)). This can be explained as bigger numbers of recommended tags imply lower precision values because more non-relevant tags are recommended. Nevertheless, it is interesting to observe that the increase in F of Cla over Gen is specially notorious for large numbers of recommended tags ($|\Gamma_\mathrm{R}| > 8$, Fig. 6(b)). This highlights the superiority of Cla over Gen when larger number of tags are recommended, and suggests that Cla is able to provide more comprehensive and relevant recommendations.

6. Conclusion and discussion

In this work we describe and evaluate two tag recommendation methods in the audio clip sharing context of Freesound. One general tag recommendation method (Gen) was introduced in previous work by the authors. The other method, which is class-based (CLA), is an original contribution of this article. It extends the former in two main aspects: it automatically determines to which class an audio clip belongs and it produces specific recommendations for different audio classes. As both tag recommendation methods (GEN and CLA) are folksonomy-based, they are easily generalisable to other multimedia domains. However, the CLA method requires the definition of K classes of resources in the particular domain, and the building of a ground truth to train the classifier needed to perform recommendations. The main bottleneck in terms of scalability lies in the computation of the tag-tag similarity matrices that inform the candidate selection step. However, these matrices can be computed offline, and their size can be easily reduced by raising the threshold ω during the construction of the association matrix. This will discard those tags whose frequency of occurrence is below that threshold (Sec. 2.1). That means that our recommendation methods can scale well to even bigger amounts of data, as the number of tags above the threshold ω will grow much more slowly than the number of resources.

A limitation of the proposed recommendation methods is that they can suffer the *cold-start* problem if deployed to collaborative tagging systems which have not enough data to derive reliable tag-tag similarity matrices. Although our recommendation methods have not been designed for collaborative tagging systems with scarce data, it would be interesting to evaluate how fast the methods could acquire enough data from user annotations in order to provide meaningful recommendations. In other words, it would be interesting to investigate how big the folksonomy of a collaborative tagging system should be to enable our tag recommendation methods to provide meaningful recommendations. We hypothesise that, on a first step of the implementation of the system, tag-tag similarity matrices would need to be recomputed very often as relatively small changes in the folksonomy could have a big impact on the resulting similarity matrix. In that case, the system would quickly learn from user tagging behaviour and recommendations would quickly start to become more diverse. Besides the similarity matrices, the CLA method also needs annotation data to train the classifier. However, a collaborative tagging system could start using the GEN method until enough data would be collected to build the ground truth and train the classifier.

As a second contribution, we perform a user-based evaluation through an online experiment. In it, participants had to annotate several audio clips with the help of the different tag recommendation strategies. We logged the activity of the participants and analysed these logs with the goal of comparing the considered methods and, in addition, getting more insight into the positive and negative aspects of tag recommendation systems in general. To the best of our knowledge, this is one of the very few user-based evaluations carried out for a tag recommendation task. Finally, as a further contribution, we complement the user-based evaluation with a prediction-based evaluation, following a well-established methodology and not considering any user input.

As a main result, we have seen that class-based recommendation reports statistically significantly better scores than general recommendation, both in the user-based and prediction-based evaluations. The difference in scoring is, in absolute terms, more prominent for the user-based evaluation. Moreover, it further improves when considering only expert users. This suggests that the class-based method does indeed bring some improvements in the recommendations compared to general recommendation, and that these improvements are more noticeable to expert users.

Among all annotations that participants performed during the online experiment, approximately one third of them correspond to tags recommended by the system (for both GEN and CLA methods). That by itself brings evidence with regard to the general utility of tag recommendation systems. However, the found results also indicate that tag suggestions referring to generic concepts or sound classes tend to be more useful than recommendations of very concrete tags describing specific sound characteristics. Participants found tag suggestions more useful for sounds under SOUNDSCAPE and SPEECH categories. We hypothesise that this happens because these categories are more suited to the use of generic tags. MUSIC and SAMPLE audio classes require of annotations describing very specific musical concepts such as pitch, tonality or beats per minute. Participants had difficulties in annotating such concepts, as they are problematic to annotate without having a certain knowledge of the recording context (i.e., without being the author of the audio clip) and because tag recommenders tend to produce less meaningful suggestions in these cases. All these often overlooked qualitative evaluation aspects also represent a valuable contribution of the present article.

We believe that, in order to build better tag recommendation systems, those should be more aware of the particular contexts of the resources being described and should extensively exploit all available knowledge. To generate tag suggestions describing more concrete aspects of sound characteristics we need systems that know about the specifics of the audio domain, such as which are the most relevant properties of audio clips for different audio categories, and how to automatically estimate some of these properties. For that reason, we believe that future tag recommenders should take advantage of knowledge representation mechanisms such as ontologies to be able to include tags describing the audio domain in some structured representation, and to be able to produce informed recommendations based on reasoning and users' input. Such a system should contribute in greatly improving online resource descriptions and thus facilitating and providing new opportunities for content reuse.

Acknowledgements

We would like to thank Perfecto Herrera for his help in designing the online experiment and also all Freesound users that participated. This work has been supported by BES-2010-037309 FPI from the Spanish Ministry of Science and Innovation (TIN2009-14247-C02-01; F.F.), 2009-SGR-1434 from Generalitat de Catalunya (J.S.), JAEDOC069/2010 from CSIC (J.S.), ICT-2011-8-318770 from the European Commission (J.S.), and FP7-2007-2013 / ERC grant agreement 267583 (CompMusic; F.F., X.S.).

References

[1] Anderson, A., Ranghunathan, K., Vogel, A., 2008. Tagez: Flickr tag recommendation, in: Proceedings of the 23rd Conference on Artificial Intelligence (AAAI 2008).

[2] Cao, H., Xie, M., Xue, L., Liu, C., 2009. Social tag prediction based on supervised ranking model, in: Proceedings of the Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD), Discovery Challenge Workshop, pp. 35–48.

[3] Chen, Z., Cao, J., Song, Y., Guo, J., Zhang, Y., Li, J., 2010. Context-oriented web video tag recommendation, in: Proceedings of the 19th International Conference on World Wide Web (WWW 2010), p. 1079.

[4] De Meo, P., Quattrone, G., Ursino, D., 2009. Exploitation of semantic relationships and hierarchical data structures to support a user in his annotation and browsing activities in folksonomies. Journal of Information Systems 34, 511–535.

[5] Font, F., Roma, G., Serra, X., 2013a. Freesound Technical Demo, in: Proceedings of the 21st ACM Conference on Multimedia (ACM MM 13), pp. 411–412.

[6] Font, F., Serrà, J., Serra, X., 2013b. Audio clip classification using social tags and the effect of tag expansion, in: Proceedings of the 53rd AES Conference on Semantic Audio.

[7] Font, F., Serrà, J., Serra, X., 2013c. Folksonomy-based tag recommendation for collaborative tagging systems. International Journal on Semantic Web and Information Systems 9, 1–30.

[8] Garg, N., Weber, I., 2008. Personalized, interactive tag recommendation for flickr, in: Proceedings of the 2nd ACM Conference Recommender systems (RecSys 08), pp. 67–74.

[9] Halpin, H., Robu, V., Shepard, H., 2006. The dynamics and semantics of collaborative tagging, in: Proceedings of the 1st Semantic Authoring and Annotation Workshop, pp. 1–21.

[10] Hogg, R.V., Craig, A.T., 1995. Introduction to Mathematical Statistics. 5th ed., Prentice Hall.

[11] Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6, 65–70.

[12] Ivanov, I., Vajda, P., Goldmann, L., Lee, J.S., Ebrahimi, T., 2010. Object-based tag propagation for semi-automatic annotation of images, in: Proceedings of the International Conference on Multimedia Information Retrieval, pp. 497–506.

[13] Jäschke, R., Eisterlehner, F., Hotho, A., Stumme, G., 2009. Testing and evaluating tag recommenders in a live system, in: Proceedeings of the 3rd ACM Conference on Recommender systems (RecSys 09), pp. 369–372.

[14] Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., Stumme, G., 2007. Tag Recommendations in Folksonomies, in: Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, pp. 506–514.

[15] Li, J., Wang, J.Z., 2006. Real-time computerized annotation of picture, in: Proceedings of the 14th ACM Multimedia Conference, pp. 911–920.

[16] Lipczak, M., 2008. Tag recommendation for folksonomies oriented towards individual users, in: Proceedings of the Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD), Discovery Challenge Workshop, pp. 84–95.

[17] Mann, H., Whitney, D., 1947. On a test of whether one of two random variables is stochastically larger than the other. The annals of mathematical statistics 18, 50–60.

[18] Marinho, L.B., Preisach, C., Schmidt-Thieme, L., 2009. Relational classification for personalized tag recommendation, in: Proceedings of the Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD), Discovery Challenge Workshop, pp. 7–15.

[19] Naaman, M., Nair, R., 2008. ZoneTags Collaborative Tag Suggestions: What is This Person Doing in My Phone? IEEE MultiMedia 15, 34–40.

[20] Rendle, S., Schmidt-Thieme, L., 2009. Factor models for tag recommendation in bibsonomy, in: Proceedings of the Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD), Discovery Challenge Workshop, pp. 235–242.

[21] Salzberg, S.L., 1997. On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach. Journal of Data Mining and Knowledge Discovery 1, 317–328.

[22] Sigurbjörnsson, B., Zwol, R., 2008. Flickr tag recommendation based on collective knowledge, in: Proceedings of the 17th International Conference on World Wide Web (WWW 2008), pp. 327–336.

[23] Sood, S.C., Owsley, S.H., Hammond, K.J., Birnbaum, L., 2007. TagAssist: Automatic Tag Suggestion for Blog Posts, in: Proceedings of the 1st International Conference on Weblogs and Social Media (ICWSM 07), pp. 1–8.

[24] Sordo, M., 2012. Semantic Annotation of Music Collections: A Computational Approach. Ph.D. thesis. Universitat Pompeu Fabra, Barcelona.

[25] Toderici, G., Aradhye, H., Pasca, M., Sbaiz, L., Yagnik, J., 2010. Finding meaning on youtube: Tag recommendation and category discovery, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010), pp. 3447–3454.

[26] Turnbull, D., Barrington, L., Torres, D., Lanckriet, G., 2008. Semantic Annotation and Retrieval of Music and Sound Effects. IEEE Transactions On Audio Speech And Language Processing 16, 467–476.