International Journal on Semantic Web and Information Systems

Folksonomy-based tag recommendation for collaborative tagging systems

Frederic Font
Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain

Joan Serrà
Artificial Intelligence Research Institute (IIIA-CSIC), Spanish National Research Council, Bellaterra, Spain

Xavier Serra
Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain

Abstract

Collaborative tagging has emerged as a common solution for labelling and organising online digital content. However, collaborative tagging systems typically suffer from a number of issues such as tag scarcity or ambiguous labelling. As a result, the organisation and browsing of tagged content is far from being optimal. In this work we present a general scheme for building a folksonomy-based tag recommendation system to help users tagging online content resources. Based on this general scheme, we describe eight tag recommendation methods and extensively evaluate them with data coming from two real-world large-scale datasets of tagged images and sound clips. Our results show that the proposed methods can effectively recommend relevant tags, given a set of input tags and tag co-occurrence information. Moreover, we show how novel strategies for selecting the appropriate number of tags to be recommended can significantly improve methods performances. Approaches such as the one presented here can be useful to obtain more comprehensive and coherent descriptions of tagged resources, thus allowing a better organisation, browsing and reuse of online content. Moreover, they can increase the value of folksonomies as reliable sources for knowledge-mining.

Keywords: Intelligent System; Recommendation System; Collaborative Tagging; Folksonomies

Introduction

Collaborative tagging has emerged as a common and successful solution for labelling and organising huge amounts of digital content, being adopted by many well-known sites such as Youtube, Flickr, Last.fm, or Delicious (Marlow, Naaman, Boyd, & Davis, 2006). In collaborative tagging, users assign a number of free-form semantically-meaningful textual labels (tags) to information resources. These tags can be then used for many purposes, including retrieval, browsing and categorisation (Bischoff, Firan, Nejdl, & Paiu, 2008). For instance, they can be used for matching user queries with resources tags, or for building tag clouds to navigate across resources. Such usages are of special importance for platforms that share multimedia content such as videos, images, or audio, since such contents can not be so directly and straightforwardly indexed as it would be done with textual data like books or web pages (Bischoff et al., 2008). Because of this importance, collaborative tagging systems have been widely researched in the last few years. In particular, a focus has been given to collaborative tagging dynamics and user behaviour (Marlow et al., 2006; Halpin, Robu, & Shepard, 2006; Golder & Huberman, 2006; Farooq et al., 2007) and to automatic tag classification methods based on user motivations (Bischoff et al., 2008; Cantador, Konstas, & Jose, 2011).

Nevertheless, collaborative tagging systems suffer from a number of well-known issues (Halpin et al., 2006; Cantador et al., 2011), which include tag scarcity, the use of different labels to refer to a single concept (synonymy), the ambiguity in the meaning of certain labels (polysemy), the commonness of typographical errors, the use of user-specific naming conventions, or even the use of different languages. One strategy for trying to overcome these problems, and thus to obtain more comprehensive and consistent tag assignments, is the use of tag recommendation systems to help users in the tagging process (Jäschke, Marinho, Hotho, Schmidt-Thieme, & Stumme, 2007). In that case, when users are labeling online resources, tag recommendation systems automatically suggest new tags that can also be meaningful or relevant for the resource being described. This way, tag recommendation serves the purpose of consolidating the tag vocabulary among users in a collaborative tagging system (Jäschke et al., 2007). In addition, tag recommendation systems can be used, in an off-line mode, to extend the descriptions of information resources by automatically adding new tags.

Here we describe a general scheme for tag recommendation in large-scale collaborative tagging systems. Our approach is folksonomy-based, meaning that we do not perform any content analysis of the information resources for which we perform tag recommendations, but uniquely rely on the tag co-occurrence information that can be derived from the folksonomy itself. A particularly interesting aspect of our tag recommendation scheme is a step focused on automatically selecting the number of tags to recommend given a list of candidates with assigned scores. Other tag recommendation methods found in the literature generally do not consider this aspect and evaluate their solutions at different values of

This work is partially supported under BES-2010-037309 FPI grant from the Spanish Ministry of Science and Innovation for the TIN2009-14247-C02-01 DRIMS project. JS acknowledges 2009-SGR-1434 from Generalitat de Catalunya, JAEDOC069/2010 from Consejo Superior de Investigaciones Científicas, and FP7-ICT-2011-8-318770 from the European Commission.

K recommended tags (see Related work). Moreover, as the scheme we describe only relies on tag information derived from a folksonomy, it is rather domain-independent and could be easily adapted to other collaborative tagging systems, either alone or as a complement of more specific content-based strategies. We believe that a tag recommendation method such as the one we propose here can be useful to obtain more comprehensive and coherent descriptions of tagged resources, and help the emergence of less noisy and more consistent folksonomies. This can greatly benefit organisation, browsing and reuse of online content, and also leverage the value of folksonomies as reliable sources for knowledge-mining (Al-Khalifa & Davis, 2007; Limpens, Gandon, & Buffa, 2009).

We propose eight tag recommendation methods which are based on the aforementioned general scheme. The proposed methods, jointly with several baselines, are evaluated with data coming from Freesound[1] (an online audio clip sharing site with more than two million registered users and 150,000 sounds (Akkermans et al., 2011)) and Flickr[2] (a well known photo sharing site that, according to Wikipedia ("Flickr", 2012), has more than 50 million registered users and six billion photos). For the best scoring methods, we also analyse the impact of their configurable parameters. Overall, we have performed more than 100 experiments and computed around 7 million tag recommendations for the resources of Freesound and Flickr.

The rest of the paper is organized as follows. First we comment on other tag recommendation approaches found in related work. Then we describe the different steps of our tag recommendation scheme and the strategies we propose to compute each step. After that we outline the characteristics of the evaluation datasets and describe the methodology we followed to evaluate our methods and the baselines. Results are then reported first in general terms of accuracy and number of recommended tags and then focusing on particular parameters of the recommendation methods. We conclude the paper with a discussion about our findings and future work.

Related work

A wide variety of tag recommendation methods for images and audio clips can be found in the literature. In general, similar strategies have been applied in both fields. On the one hand, typical approaches are based on the extraction of features from image/audio content and their posterior analysis. In (Farooq et al., 2007) and (Ivanov, Vajda, Goldmann, Lee, & Ebrahimi, 2010), the authors propose a method for propagating the tags of an annotated image to other images which are considered to be similar according to some content-based similarity metric. Similar approaches have been applied to music (Sordo, 2012) and audio clips (Martínez, Celma, Sordo, Jong, & Serra, 2009). A slightly more complex approach, also based on content analysis, is the use of machine learning techniques to learn mappings between tags and image/audio low-level features. In that direction, we find relevant work in (Barrington, Chan, Turnbull, & Lanckriet, 2007; Turnbull, Barrington, Torres, & Lanckriet, 2008) for the audio case, and in (Barnard et al., 2003; Li & Wang, 2006) for the image case. Due to the content-based nature of these strategies, they are not directly comparable to the approach we propose here.

[1] `www.freesound.org`.
[2] `www.flickr.com`.

On the other hand, methods for tag recommendation have also been proposed which are based on folksonomies, that is to say, considering user-tag-resource relations from a collaborative tagging system. Sigurbjörnsson and Zwol (2008) and Garg and Weber (2008) propose methods for image tag recommendation which are based on tag co-occurrences derived from the Flickr folksonomy. The general idea is that a tag-tag similarity matrix can be derived by considering tags with high co-occurrence as similar tags (Markines et al., 2009). Recommendations are then performed on the basis of such similarity matrix. In (Meo, Quattrone, & Ursino, 2009) and (Jäschke et al., 2007), more complex strategies for tag recommendation based on folksonomies are described and evaluated with data from Delicious, BibSonony, and Last.fm[3]. They use hierarchical tag structures and the FolkRank (Hotho, Jäschke, Schmitz, & Stumme, 2006) ranking algorithm, respectively. There are also some tag recommendation approaches which mix content and folksonomy-based strategies such as (Wu, Yang, Yu, & Hua, 2009), where three different lists of candidate tags are constructed taking into account different tag-tag similarity measures (based on image content and tag co-occurrence) and then non-linearly combined to provide a final ranking. Noticeably, all the aforementioned approaches produce a list of sorted candidates with scores, with no further indication of how many of them should be recommended.

A different approach in tag recommendation methods relates to the personalisation of the recommendations for particular user tagging styles. Jäschke et al. (2007) approach such personalised recommendation by using collaborative filtering techniques. Garg and Weber (2008) propose a variant of their recommendation method that collects candidate tags among previously used tags from the same user. In (Lipczak, 2008), user profiles are build as sets of previously used tags and these are promoted in the recommendation process. Other approaches take advantage of probabilistic and machine learning techniques to learn latent interactions between users, resources and tags (Rendle & Schmidt-Thieme, 2009; Marinho, Preisach, & Schmidt-Thieme, 2009; Cao, Xie, Xue, & Liu, 2009). Personalised tag recommenders are specially suited for collaborative tagging systems where the same resources are tagged by different users (e.g., Delicious, BibSonomy, Last.fm; so called *broad* folksonomies (Wal, 2005)). In these systems, users annotate resources for self-organisation purposes and, therefore, it is useful to tailor recommendations to the tagging style of each user. However, in other systems resources are only annotated by their original contributors (e.g., Flickr, Freesound; called *narrow* folksonomies), and indexing is only performed on the basis of these annotations. In such cases, a common tagging style across users is preferred so that resources are tagged more uniformly and can be better organised (Lipczak, 2008). The tag recommendation method we propose in this work is oriented toward narrow folksonomies.

The majority of the tag recommendation methods cited above are evaluated in terms of precision and recall at specific numbers of K recommended tags. Thus, the output of the recommendation task is a sorted list of scored candidate tags without any specified length. There are, however, a few methods in which the number of recommended tags is automatically determined following some simple rules. In (Marinho et al., 2009), where recommendations are personalised to particular users, the length of the recommendation is limited to the average number of tags per resource that a given user provided in the

[3] `www.delicious.com`, `www.bibsonomy.org`, `www.last.fm`

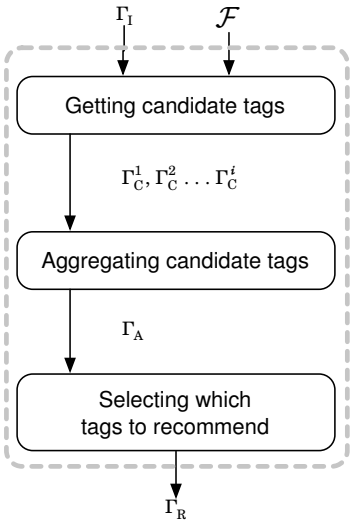

Figure 1. Block diagram of the described tag recommendation scheme.

past. Cao et al. (2009) propose a method for ranking and sorting candidate tags, and systematically recommend half of the candidates. The approach described in (Rendle & Schmidt-Thieme, 2009) limits recommendations using simple heuristics based on statistics of the folksonomy such as the average number of tags per resource. Here we explore more in depth the problem of automatically determining the number of tags to recommend and propose novel strategies based on scores assigned to candidate tags.

Folksonomy-based tag recommendation

In this work we describe a general scheme for tag recommendation that, given a set of input tags Γ_I and a folksonomy $\mathcal{F}$, outputs a set of recommended tags Γ_R (Fig. 1). The folksonomy can be defined as a set of tag assignments $\mathcal{F} \subseteq U \times T \times R$, where U, T, and R denote sets of users, tags, and resources, respectively (Mika, 2007)[4]. The described scheme is composed of three independent steps: 1) Getting candidate tags, 2) Aggregating candidate tags, and 3) Selecting which tags to recommend. For Step 1, we propose three variants based on different similarity measures widely used in the literature (tag co-occurrence, cosine and Jaccard similarity, (Halpin et al., 2006; Jäschke et al., 2007; Mika, 2007; Sigurbjörnsson & Zwol, 2008; Meo et al., 2009; Markines et al., 2009)). For Step 2, we propose two aggregation strategies (Similarity-based and Rank-based). For Step 3, we propose four selection strategies (Percentage, Statistical test, Kernel percentage and Linear regression). What follows is a brief overview of these steps. In-depth descriptions are given in subsequent sections.

• Step 1: Getting candidate tags. Given Γ_I and $\mathcal{F}$, this step retrieves a set of N candidate tags Γ_C^i for each input tag Γ_{I_i}. The retrieval of these candidates is based on tag-tag semantic similarity measures derived from $\mathcal{F}$.

[4]Mika (2007) uses the terminology Actor, Concept, and Instance (A, C and I) to denote what we call User, Tag, and Resource (U, T and R). We adopted the latter terminology as it more closely relates with the data we are dealing with.

• Step 2: Aggregating candidate tags. This step takes the sets of candidates Γ_{C}^{i}, assigns a score value to each individual tag, and aggregates all candidates to form a single list of tags with assigned scores Γ_{A}.

• Step 3: Selecting which tags to recommend. This step automatically selects which tags to recommend given the candidate tags and score values of Γ_{A}. The output of this step is the final recommendation Γ_{R}.

Step 1: Getting candidate tags

We start the recommendation process by obtaining a number of related candidate tags to the set of input tags Γ_{I}. For each input tag Γ_{I_i}, we get a set of candidates Γ_{C}^{i} by selecting the N closest tags to Γ_{I_i} according to a tag-tag semantic similarity measure. For this purpose, we build a tag-tag similarity matrix $\mathcal{S}$ based on the tag assignment information contained in the folksonomy $\mathcal{F}$. Notice that $\mathcal{S}$ is not dependent of the particular Γ_{I_i} for which we are selecting candidates. Therefore, it only needs to be computed once for a given $\mathcal{F}$[5].

The folksonomy $\mathcal{F}$ can be naturally represented as a tripartite hypergraph $\mathcal{G}(\mathcal{F}) = \langle V, E\rangle$, where vertices are given by three finite sets of objects, $V = U \cup T \cup R$, and each edge E represents a tag-resource association performed by a user, $E = \{\{u, t, r\}|(u, t, r) \in \mathcal{F}\}$ (Mika, 2007). We unfold $\mathcal{G}(F)$ into the bipartite graph $\mathcal{TR}$, which only reflects the associations between tags and resources. We represent the bipartite graph $\mathcal{TR}$ as a matrix $\mathcal{D} = \{d_{i,j}\}$, where $d_{i,j} = 1$ if tag t_i has been used to label resource r_j, and $d_{i,j} = 0$ otherwise. We then define the matrix

$$\mathcal{S} = \mathcal{D}\mathcal{D}', \tag{1}$$

which corresponds to a one-mode network connecting tags on the basis of shared resources (Mika, 2007) (the symbol $'$ denotes matrix transposition). Elements $s_{i,j}$ of $\mathcal{S}$ indicate the number of resources in which tags t_i and t_j appear together. Therefore, the diagonal of $\mathcal{S}$ represents the total number of different resources labeled with a tag $t_{i=j}$.

At this point, $\mathcal{S}$ can be interpreted as a tag-tag similarity matrix based on absolute co-occurrence. That is to say, the similarity between tags t_i and t_j is represented by the total number of times they appear together. This is the first similarity measure we use for our tag recommendation method. In order to obtain the rest of aforementioned similarity measures, we have to apply different normalisation procedures to $\mathcal{S}$. Cosine similarity can be obtained as

$$s_{t_i,t_j} = \frac{\sum_n d_{i,n} d_{j,n}}{\sqrt{\sum_n {d_{i,n}}^2}\sqrt{\sum_n {d_{j,n}}^2}}. \tag{2}$$

Given that rows $\mathbf{d}_i$ and $\mathbf{d}_j$ are bit vectors (the only possible values are 0 or 1), $\sum_n d_{i,n} d_{j,n}$ is equivalent to the absolute co-occurrence between tags t_i and t_j, while $\sum_n {d_{i,n}}^2$ and $\sum_n {d_{j,n}}^2$ is equivalent to the total number of occurrences of tags t_i and t_j, respectively (the total number of resources labeled with t_i and t_j, respectively). Therefore, cosine similarity is equivalent to dividing each element in $\mathcal{S}$ (Eq. 1) by $\sqrt{s_{t_i,t_i}}\sqrt{s_{t_j,t_j}}$. In a similar way, the

[5] As it is described later in the Datasets subsection, we filter out the least frequent tags of our folksonomy in order to reduce the computational complexity of $\mathcal{S}$.

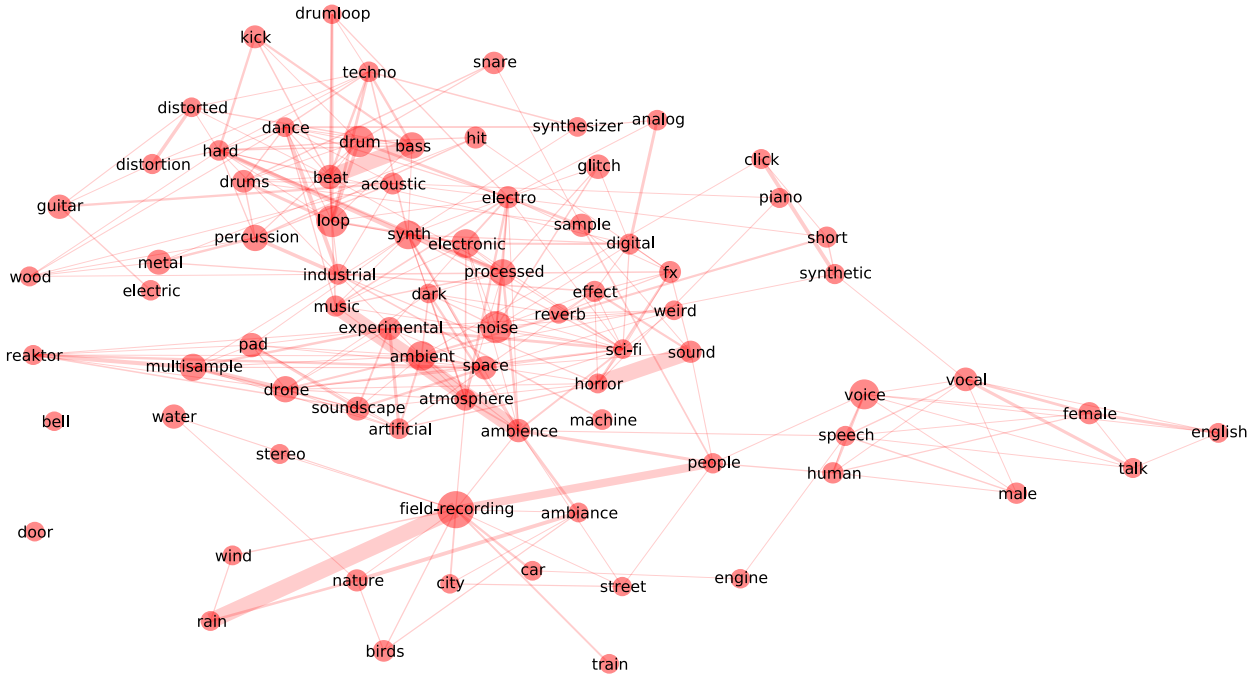

Figure 2. Graph visualisation of a tag-tag similarity matrix $\mathcal{S}$ built using cosine similarity and the Freesound folksonomy. Edge widths represent the cosine similarity between two tags. Tag size is a logarithmic function of the absolute tag frequency. For visualisation purposes, only edges above a certain degree of similarity and tags above a certain level of absolute frequency are shown.

Jaccard index can be obtained as

$$s_{t_i,t_j} = \frac{\sum_n d_{i,n} d_{j,n}}{\sum_n {d_{i,n}}^2 + \sum_n {d_{j,n}}^2 - \sum_n d_{i,n} d_{j,n}}. \tag{3}$$

Hence, the Jaccard index is equivalent to dividing each element in $\mathcal{S}$ (Eq. 1) by $s_{t_i,t_i} + s_{t_j,t_j} - s_{t_i,t_j}$. Independently of the used similarity measure, $\mathcal{S}$ can be represented as a graph where nodes correspond to tags and edges represent the similarities between two tags (i.e., s_{t_i,t_i}, see Fig. 2).

Once we have a tag similarity matrix $\mathcal{S}$, we iterate over the input tags Γ_{I} and get, for each element Γ_{I_i}, a set of candidates Γ^i_{C}. Specifically, we select the N most similar tags to Γ_{I_i} (i.e., the N most similar graph neighbours of Γ_{I_i}) and keep these similarity values for further processing. Hence, for instance, if our method is fed with three input tags, it will get a maximum of $3N$ candidate tags (separated into 3 sets), provided that all three input tags have at least N graph neighbours.

Step 2: Aggregating candidate tags

The next step of our tag recommendation scheme takes all the sets of candidates Γ^i_{C}, assigns a score value ϕ_j to every candidate $\Gamma^i_{\mathrm{C}_j}$ in Γ^i_{C}, and then aggregates all sets into a single list of tags with assigned scores Γ_{A}. The output of this step, Γ_{A}, is a list of tuples

where each element contains a tag and its assigned score. To accomplish this step, we propose two different strategies:

Similarity-based strategy. In the Similarity-based strategy, the j-th candidate tag $\Gamma^i_{\mathrm{C}_j}$ of Γ^i_{C} is assigned a score ϕ_j that directly corresponds to the similarity value between the candidate tag and the corresponding input tag Γ_{I_i}, i.e., $\phi_j = s_{u,v}$, where $u = \Gamma^i_{\mathrm{C}_j}$ and $v = \Gamma_{\mathrm{I}_i}$. After that, the list of tuples Γ_{A} is constructed as the union of all sets of candidates Γ^i_{C} and their scores. If a particular tag has duplicates in Γ_{A} (which can happen if a given tag appears in several sets of candidates Γ^i_{C}), we only keep one occurrence and set its score to the sum of all the scores of the duplicates of that tag. This way we promote tags that are considered to be similar to more than one input tag. As we do not want to recommend tags that are already part of Γ_{I}, we remove any occurrences of these tags in Γ_{A}. We finally normalise the assigned scores by dividing them by the number of input tags $|\Gamma_{\mathrm{I}}|$.

Rank-based strategy. The Rank-based strategy only differs from the Similarity-based strategy above in the way scores are assigned. Instead of directly using the similarity values from Step 1, we assign discrete ranks. For this purpose, we sort each set Γ^i_{C} by similarity values in descending order, and assign scores as $\phi_j = N - (r - 1)$, where r is the position of the j-th tag in Γ^i_{C} after sorting (thus r ranges from 1 to N). Notice that the most similar tag to every input tag will be assigned a score of N. Even if a particular set Γ^i_{C} contains less than N tags (meaning that corresponding input tag Γ_{I_i} has less than N neighbours in the graph representation of $\mathcal{S}$), the score we assign to the most similar tag will be of N. After score assignment we proceed exactly as with Similarity-based aggregation: constructing Γ_{A} as the union of all sets Γ^i_{C}, merging duplicate tags in Γ_{A} by adding their scores, removing tags appearing in Γ_{I}, and normalising score values by $|\Gamma_{\mathrm{I}}|$. An example comparing the two aggregation strategies is shown in Table 1.

Step 3: Selecting which tags to recommend

Once we have computed Γ_{A}, we select which of these tags should be recommended. For that we consider four strategies that take into account the scores ϕ of Γ_{A} to automatically determine a threshold ε. The set of recommended tags Γ_{R} is then formed by all the elements of Γ_{A} whose scores are equal or above ε.

Percentage strategy. This is a straightforward strategy where ε is determined as a percentage of the highest score in Γ_{A} by

$$\varepsilon = (1 - \alpha) \cdot \max(\phi),$$

where α is the percentage parameter that must be configured. Following the example shown in Table 1, and taking $\alpha = 0.05$, only one tag would be recommended for the Similarity-based aggregation ($\varepsilon = (1 - 0.05) \cdot 0.307 = 0.292$; $\Gamma_{\mathrm{R}} = \{\texttt{birds}\}$) and three tags would be recommended for the Rank-based aggregation ($\varepsilon = (1 - 0.05) \cdot 100 = 95$; $\Gamma_{\mathrm{R}} = \{\texttt{birds}$, `ambiance`, `south-spain`$\}$).

Kernel percentage strategy. The Kernel percentage strategy has two steps. First, we estimate the probability density function $\mathcal{P}$ of ϕ, the scores of Γ_{A}. For that purpose, we use a kernel density estimator (Silvermann, 1986), a fundamental data smoothing technique.

Table 1: Example of the output of the aggregation step using the Freesound folksonomy with $\Gamma_{\mathrm{I}} = \{\texttt{field-recording}, \texttt{nature}\}$ (for both Similarity-based and Rank-based strategies, $N = 100$). Candidate tags are sorted by their score values. A score of 100 for the tag `birds` in the Rank-based aggregation means that it is the most similar tag to both `field-recording` and `nature` ($100/2+100/2 = 100$). Notice that due to the use of different scoring methods, Similarity-based and Rank-based aggregation strategies produce different sorting of candidate tags and score distributions.

	Γ_{A}			
	Similarity-based		**Rank-based**	
#	**Tag**	ϕ	**Tag**	ϕ
1	`birds`	0.307	`birds`	100.0
2	`south-spain`	0.244	`ambiance`	97.0
3	`ambiance`	0.229	`south-spain`	97.0
4	`spring`	0.180	`summer`	92.0
5	`summer`	0.169	`spring`	91.5
6	`bird`	0.162	`bird`	90.0
7	`insects`	0.157	`thunder`	82.5
8	`donana`	0.155	`rain`	82.0
9	`ambience`	0.151	`ambience`	80.0
10	`forest`	0.147	`forest`	79.5
11	`thunder`	0.145	`weather`	79.5
12	`rain`	0.139	`field`	79.0
13	`marshes`	0.139	`water`	77.5
14	`weather`	0.137	`birdsong`	75.5
15	`water`	0.129	`purist`	75.5
16	`purist`	0.129	`donana`	72.5
17	`field`	0.127	`street-noise`	71.5
18	`birdsong`	0.127	`insects`	71.5
19	`street-noise`	0.121	`thunderstorm`	70.0
20	`atmos`	0.118	`storm`	70.0
	+ 186 more			

The bandwidth of the kernel is automatically determined using Scott's Rule (Scott, 2008). Then, the threshold is defined as the ε that satisfies

$$\int_{\min(\phi)}^{\varepsilon} \mathcal{P}(\phi)\,\mathrm{d}\phi = (1-\beta)\int_{\min(\phi)}^{\max(\phi)} \mathcal{P}(\phi)\,\mathrm{d}\phi, \tag{4}$$

where β is the percentage parameter that must be configured. Therefore, β determines a percentage of the area of $\mathcal{P}$ which we consider to include suitable tags for the recommendation (Fig. 3). The bigger the parameter β, the smaller the threshold ε becomes and thus the more tags are finally recommended.

The idea behind this strategy is that, understanding the scores of Γ_{A} as a sample extracted from a population of scores with an underlying distribution $\mathcal{P}$, the threshold ε will be better determined considering a percentage of the area of that underlying distribution rather than as a percentage of the maximum observed score (as we propose in the Percentage strategy).

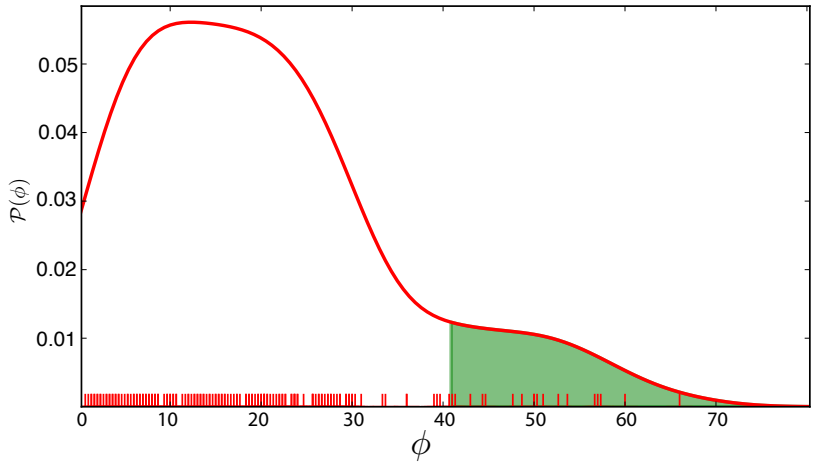

Figure 3. Example of the Kernel percentage strategy for selecting which tags to recommend (using $\beta = 0.05$). The curve represents the estimated $\mathcal{P}$ of the scores of Γ_A. Vertical markers on the x-axis show the actual positions of candidate tag scores. The shaded zone in the right of the figure corresponds to the 5% of the total area of $\mathcal{P}$. Recommended tags are those under that zone.

Statistical test strategy. Here we also estimate the probability density function $\mathcal{P}$ of ϕ using a kernel density estimator as before. However, to determine the threshold ε we start an iterative process where, in each iteration, we select a slice of $\mathcal{P}$ and perform a statistical test for normality according to

$$\mathrm{AD}(\mathcal{P}_{\varepsilon:\max(\phi)}), \tag{5}$$

where the function AD is the Anderson-Darling test for normality (Scholz & Stephens, 1987), and $\mathcal{P}_{\varepsilon:\max(\phi)}$ is a slice of $\mathcal{P}$ that goes from ε to $\max(\phi)$. In each iteration, ε takes a different value such that

$$\varepsilon = \max(\phi) - i \cdot \frac{\max(\phi) - \mathrm{mix}(\phi)}{100}, \tag{6}$$

where i is the number of the current iteration ($i \in 1, 2, 3, ...$). We stop the iterative process when the test fails for the first time (i.e., when the probability of having an independent Gaussian distribution is not statistically significant). The final threshold takes the value of ε at that iteration (Fig. 4).

The idea behind this process is that for a given set of candidate tags there will be a subset of good tags for the recommendation exhibiting a normal, independent distribution separated from the rest of candidates. The statistical test fails when it detects departures from normality and, according to our hypothesis, this will happen when non-meaningful candidate tags start affecting $\mathcal{P}$. Notice that this strategy, in practice, can be considered parameter-free as, by using the aforementioned Scott's rule, it only requires a statistical significance level from which to reject the null hypothesis of a Gaussian distribution. We here follow common practice (Scholz & Stephens, 1987) and take this significance level at 0.01. Using another common statistical significance level such as 0.05 would presumably result in less restrictive statistical tests yielding bigger sets of recommended tags.

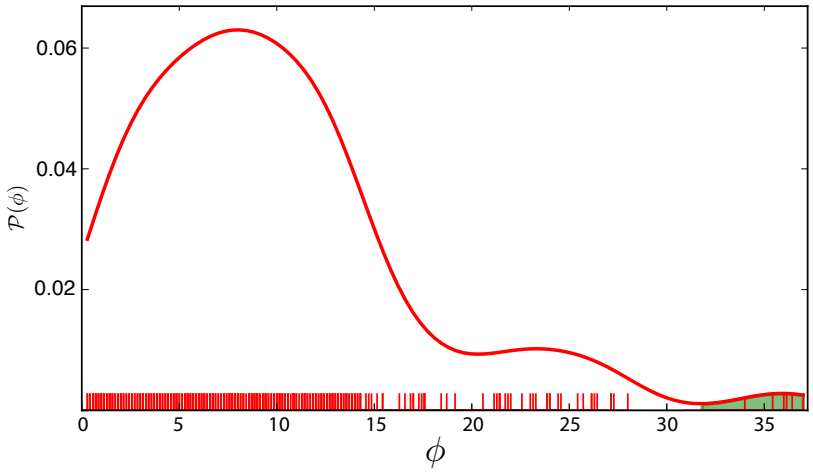

Figure 4. Example of the Statistical test strategy for selecting which tags to recommend. The curve represents the estimated $\mathcal{P}$ of the scores of Γ_{A}. Vertical markers on the x-axis show the actual positions of candidate tag scores. Recommended tags are those under the shaded zone in the right. In this example, the obtained threshold is $\varepsilon \approx 32$. Looking at the figure, it can be easily intuited that lower values of ε would cause the statistical test of Eq. 5 to fail.

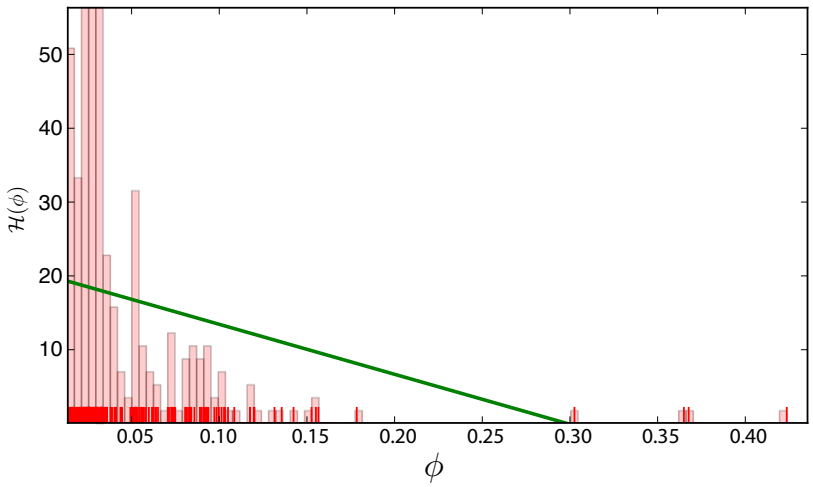

Figure 5. Example of the Linear regression strategy for selecting which tags to recommend. The straight line shows the linear regression of the histogram $\mathcal{H}$ of the scores of Γ_{A}. Vertical markers on the x-axis show the actual positions of candidate tag scores. In this example, the obtained threshold is $\varepsilon \approx 0.29$, which is the point where the linear regression crosses the y-axis. Recommended tags are those placed above 0.29.

Table 2: Basic statistics of the FREESOUND and FLICKR1M folksonomies. We see that the datasets feature comparable numbers. †Some of these tags are not semantically unique, and may include synonyms and typographic errors. ‡Users that have contributed with at least one resource.

	Before filtering		After filtering	
	Freesound	**Flickr1M**	**Freesound**	**Flickr1M**
Number of resources	118,629	107,617	118,629	107,617
Number of unique tags†	33,790	27,969	6,232	5,760
Number of contributor users‡	5,523	5,463	5,523	5,463
Number of tag assignments	782,526	927,473	730,417	882,616

Linear regression strategy. The last strategy we propose consists in calculating the least-squares linear regression of the histogram $\mathcal{H}$ of ϕ. The threshold is set at the point where the linear regression crosses the y-axis. The idea behind the Linear regression strategy is that, for a given $\mathcal{H}(\phi)$, there will be a big concentration of candidate tags with low scores, and some outliers with bigger scores that will be separated from the rest (the most suitable tags for the recommendation). Thus, the linear regression will result in a straight line with a negative slope which will be useful to separate between both groups at the point where it crosses the y-axis (Fig. 5). The more the concentration of low-scored candidates with respect to the outliers, the more pronounced the straight line will be, and the clearer the separation between both groups. Notice that this strategy is also parameter-free.

Evaluation strategy

From the combination of the different strategies above we define several tag recommendation methods which we evaluate through a tag prediction task. Essentially, what we do is to remove some tags from the resources of our datasets and then try to automatically predict them. In this section we describe the datasets and the methodology we have used for that evaluation.

Datasets

We use two real-world datasets (Table 2) collected from the collaborative tagging systems of Freesound (sound annotations) and Flickr (image annotations). In the case of Freesound, we consider all user annotations between April 2005 and September 2011, directly extracted from an anonymised version of the Freesound database[6]. From now on, we will refer to this dataset as FREESOUND. The Flickr data we use is a subset of photos taken in Barcelona, with user annotations performed approximately between 2004 and 2009. Flickr data was collected by Papadopoulos et al. (2010) and provided to us by the authors[7]. To avoid confusion with the totality of the Flickr content, we will refer to the analysed Flickr subset as FLICKR1M.

[6]Annotation data from Freesound can be obtained using the public Freesound API. Documentation can be found at `www.freesound.org/docs/api`.

[7]We assume that this data can also be provided to other researchers by requesting the original authors.

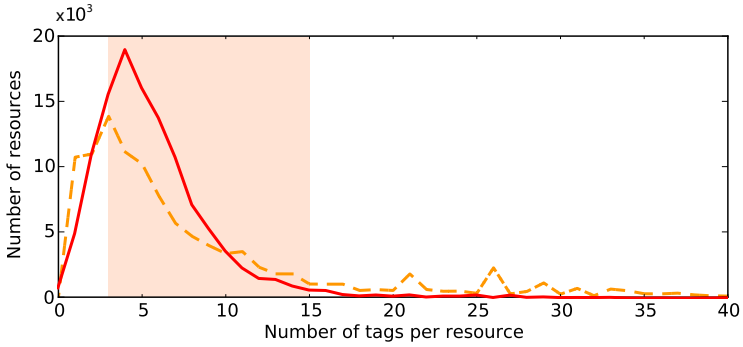

Figure 6. Distribution of number of tags per resource in FREESOUND (continuous line) and FLICKR1M (dashed line). The average number of tags per resource is 6.53 (6.47) and 7.50 (8.61) for FREESOUND and FLICKR1M, respectively (standard deviation in parenthesis).

Both Freesound and Flickr have similar uploading processes in which users first provide the content (sounds and images, respectively) and then add as many tags as they feel appropriate to each resource[8]. As opposite to other well-studied collaborative tagging systems such as Delicious or CiteULike, Freesound and Flickr feature a narrow folksonomy, meaning that resource annotations are shared among all users and therefore one single tag can only be assigned once to a particular resource (e.g., the tag `forest` can not be added twice to the same resource). Hence, we can not weight the association between a particular tag and a resource by the number of times the same association has been performed by different users. The distribution of number of tags per resource is qualitatively similar for the two datasets (Fig. 6).

We are particularly interested in recommending tags for resources that fall in the range of $[3, 15]$ tags, which are more than 80% and 65% of the total resources in FREESOUND and FLICKR1M, respectively (Fig. 6; shadowed zone). The reason for focusing on this range is that the tag recommendation scheme we propose takes as input the tags that have already been assigned to a resource. Thus, given the predictive nature of our evaluation (see below), we consider 3 tags as enough input information for our method to provide good recommendations. For resources with less than 3 tags, content-based strategies such as the ones outlined in the Related work section are probably more suited. On the other side, we intuitively consider that resources with more than 15 tags are, in general, enough well described.

Among the set of all unique tags present in FREESOUND and FLICKR1M folksonomies we apply a threshold to consider only the tags that have been used at least 10 times (i.e., tags that appear on at least 10 different resources). By this we assume that tags that have been used less than 10 times are irrelevant for our purposes. In addition, by discarding less frequent tags, we reduce the computational complexity of the calculation of $\mathcal{S}$ described in Step 1. After applying this threshold, we are left with 6,232 unique tags in the FREESOUND

[8]Since a recent software upgrade, Freesound requires a minimum of three tags to annotate a sound. However, the data we analyse is prior to the introduction of this requirement. In the case of Flickr, a single image can not be labeled with more than 75 tags, a big enough number not to be considered as a restriction for normal tagging behaviour.

folksonomy (representing $\approx 20\%$ of the total) and with 5,760 unique tags in FLICKR1M (also representing $\approx 20\%$ of the total). That also means that we filter out all tag assignments that do not associate any of these selected tags. Importantly, approximately 90% of tag assignments in both FREESOUND and FLICKR1M involve one of these tags, thus we still take into account the vast majority of the original information (Table 2).

Methodology

Our evaluation methodology follows a systematic approach based on removing a number of tags from the resources of FREESOUND and FLICKR1M and then trying to automatically predict them. The advantage of this approach is that it allows us to quickly evaluate the different recommendation algorithms without the need of human input. The main drawback is that tags that could be subjectively considered as good recommendations for a particular resource but are not present in the set of deleted tags do not count as positive results (see the discussion at the end of this article and also (Garg & Weber, 2008)).

For FREESOUND and FLICKR1M datasets separately, we perform a 10-fold cross validation following the methodology described in (Salzberg, 1997). For each fold, we build a tag similarity matrix as described in Step 1, but only using the subset of the folksonomy corresponding to the training set of resources (i.e., only considering tag assignments involving resources from the training set). For each resource in the evaluation set, we randomly delete a set of tags Γ_{D} from its originally assigned tags, yielding Γ_{I}, the input to our system. The number of tags we delete is chosen at random, with the only constraint that the length of Γ_{I} must be maintained in the range of $[3, 15]$ (see previous section). This constraint also implies that in order to be able to remove at least one tag for each resource ($|\Gamma_{\mathrm{D}}| \geq 1$), we can only consider for evaluation these resources that have at least four tags. Furthermore, we add an upper limit to the number of tags and also filter out resources with more than 16 tags. We do that to avoid outliers with many tags which would result in very low recall values. Then, we run our tag recommendation methods using the tag similarity matrix $\mathcal{S}$ derived from the training set.

As evaluation measures we compute standard precision (P), recall (R), and F-measure (F) for each individual resource according to

$$P = \frac{|\Gamma_{\mathrm{R}} \cap \Gamma_{\mathrm{D}}|}{|\Gamma_{\mathrm{R}}|} \ , \ R = \frac{|\Gamma_{\mathrm{R}} \cap \Gamma_{\mathrm{D}}|}{|\Gamma_{\mathrm{D}}|} \ , \text{ and } F = \frac{2PR}{P+R},$$

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, R and F across all resources evaluated with the particular recommendation method. Furthermore, for each individual resource we also measure the number of recommended tags $|\Gamma_{\mathrm{R}}|$. Evaluating $|\Gamma_{\mathrm{R}}|$ is important, as the longer the recommendation, the more comprehensive it probably is, and the more difficult it is to maintain high precision values (see the discussion at the end of the article). A general characterisation of the number of recommended tags per method is also obtained by averaging $|\Gamma_{\mathrm{R}}|$ across all resources evaluated with the particular recommendation method.

Table 3 summarises all tag recommendation methods we evaluate. The first group of methods (Tag recommendation methods) are the eight possible combinations of aggregation and selection strategies that we have proposed. To avoid an intractable number of possible

combinations, all methods are evaluated using only cosine similarity for Step 1, and setting $N = 100$ (getting a maximum of 100 candidates for each input tag). We choose cosine similarity as default because its widespread usage in the literature, and $N = 100$ as an intuitively big enough number of candidates per input tag. We later study the impact of the chosen similarity measure and N, using only the highest performing methods of the main evaluation. For the methods that require the configuration of a percentage parameter (SimP@α, SimKP@β, RankP@α and RankKP@β), we performed preliminary experiments with a reduced set of 10,000 resources to determine the values of α and β that reported higher average F, and only consider these values in the main evaluation.

Methods under the second group (Baseline methods) are simpler versions of the proposed methods, provided as baselines we have set up for further comparison. On the one hand, we compare with two methods that skip Step 3 and always recommend the first K tags from Γ_{A}, sorted by their scores (BRankFIX@K and BSimFIX@K). We run these algorithms for values of K ranging from 1 to 10 and report only the best accuracy. Hence, the results reported for these methods constitute an upper bound of the accuracies that can be achieved. On the other hand, we compare with an even simpler method (BRepeated@M) which, considering the union of all sets of candidates $\Gamma^1_{\mathrm{C}}, \Gamma^2_{\mathrm{C}}, \ldots\ \Gamma^i_{\mathrm{C}}$ for a given resource, only recommends tags that are repeated more than M times (independently of their scores). We run this algorithm for values of M ranging from 2 to 10 and, as above, report only the best result found.

We also compute a random baseline (BRandom) by replacing the set of Γ_{R} with a random selection (of the same length) taken from Γ_{A}. For each resource for which we recommend tags using any of the proposed methods above, we generate a random recommendation of the same length of Γ_{R}. Hence, for each proposed method, we also generate a randomised version of it. We take as the general random baseline the randomised version of all the proposed methods that reports a higher F. Notice however that these recommendations are not totally random: recommended tags are chosen from Γ_{A}, not from the set of all possible tags in Freesound or Flickr1M. Moreover, by making a recommendation of the same length as the recommendation of the non-randomised version of the method, we preserve the distribution of the number of recommended tags for each method.

Finally, methods under the third group (State of the art methods) correspond to our implementations of the tag recommendation methods described in (Garg & Weber, 2008) and (Sigurbjörnsson & Zwol, 2008), which we denote as GW and SZ, respectively. As these methods do not implement any selection step, we evaluate them for fixed values of K recommended tags ranging from 1 to 10 (and only report the best result found). In (Garg & Weber, 2008) several methods are described which contain different degrees of user personalisation. We implemented the "global" method which is not personalised and thus can be directly compared to our methods. We implemented GW and SZ following the original references and set their parameters accordingly.

Results

Recommendation accuracy

From the average P, R and F values for each one of the evaluated methods using the Freesound and Flickr1M datasets, it can be observed that Rank-based methods

Table 3: Evaluated tag recommendation methods. All methods are evaluated using cosine similarity and $N = 100$. The symbols † denote values for FREESOUND experiments and ‡ for FLICKR1M experiments.

Method	**Aggregation step**	**Selection step**
Tag recommendation methods		
SimP@α	Similarity-based	Percentage ($\alpha = 0.30$†, $\alpha = 0.20$‡)
SimST	Similarity-based	Statistical test
SimKP@β	Similarity-based	Kernel percentage ($\beta = 0.005$)
SimLR	Similarity-based	Linear regression
RankP@α	Rank-based	Percentage ($\alpha = 0.15$†, $\alpha = 0.10$‡)
RankST	Rank-based	Statistical test
RankKP@β	Rank-based	Kernel percentage ($\beta = 0.01$)
RankLR	Rank-based	Linear regression
Baseline methods		
BRankFIX@K	Rank-based	Fixed number ($K \in [1, 10]$)
BSimFIX@K	Similarity-based	Fixed number ($K \in [1, 10]$)
BRepeated@M	Repeated tags in all sets $\Gamma^{\mathrm{t}}_{\mathrm{C}}$($M \in [2, 10]$)	
BRandom	Random replacement of Γ_{R}.	
State of the art baseline methods		
GW@K	Garg & Weber (2008)	Fixed number ($K \in [1, 10]$)
SZ@K	Sigurbjörnsson & Zwol (2008)	Fixed number ($K \in [1, 10]$)

generally report higher F than Similarity-based methods (Tables 4 and 5). Comparing the F values of each Rank-based method with its Similarity-based counterpart, we observe an average increase of 0.102 and 0.049 for FREESOUND and FLICKR1M, respectively. We have tested the statistical significance of this increase by performing pairwise Kruskal-Wallis tests (Kruskal & Wallis, 1997) between the results of each Rank-based method and its Similarity-based counterpart and all have shown to be statistically significant[9], with a p-value several orders of magnitude below 0.01 (denoted as $p \ll 0.01$). These results indicate that Step 2 (Aggregating candidate tags) is better accomplished using the Rank-based strategy.

Regarding the results of the different strategies for Step 3 (Selecting which tags to recommend), we observe a very similar behaviour in FREESOUND and FLICKR1M (Tables 4 and 5, respectively). That partially supports the generalisation of the proposed strategies to different kinds of data. In both datasets, methods using the Kernel percentage strategy (either with Rank-based or Similarity-based aggregation) perform significantly worse than the others, with an average F decrease of 0.036 ($p \ll 0.01$, FREESOUND) and 0.048 ($p \ll 0.01$, FLICKR1M). Statistical test, Linear regression, and Percentage strategies report very similar F, both in FREESOUND and FLICKR1M, and specially in the case of Similarity-based aggregation. Nevertheless, the Percentage strategy in combination with Rank-based

[9]From now on, in any comparison of F we indicate the results of the statistical significance tests as the maximum of the p-values of all pairwise comparisons.

Table 4: Average precision P, recall R and F-measure F for tag recommendation methods using the FREESOUND dataset, sorted by F-measure. For the sake of readability, we only show the results of baseline methods for the values of K and M that reported higher F-measure. Baseline and state-of-the-art methods are marked in italics.

	Freesound		
Method	**Precision**	**Recall**	**F-measure**
RankP@0.15	0.444	0.532	0.437
RankST	0.443	0.537	0.433
RankLR	0.393	0.563	0.418
BRankFIX@2	*0.397*	*0.468*	*0.393*
RankKP@0.01	0.352	0.524	0.383
GW@2	*0.375*	*0.443*	*0.371*
SimLR	0.347	0.397	0.324
SimP@0.30	0.344	0.414	0.323
SimST	0.382	0.333	0.318
SimKP@0.005	0.356	0.294	0.294
BSimFIX@2	*0.303*	*0.344*	*0.293*
SZ@2	*0.286*	*0.334*	*0.281*
BRepeated@3	*0.176*	*0.678*	*0.235*
BRandom (best)	*0.006*	*0.033*	*0.011*

aggregation provides the best obtained results in both datasets, with an average F increase of 0.025 ($p \ll 0.01$, textscFreesound) and 0.039 ($p \ll 0.01$, FLICKR1M) with respect to the other selection strategies with Rank-based aggregation.

Having a look at the results of the baseline methods based on recommending a fixed number of two tags (BRankFIX@2 and BSimFIX@2) we can see that, in terms of F, they perform very similarly to the other proposed methods, and in some cases even outperform them (especially in the FLICKR1M dataset). Importantly, we have to take into account that these baseline methods only vary from our proposed methods in the last step of the recommendation process. Thus their good performance points out the effectiveness of the first two steps of the method in promoting the most relevant tags on the first positions of the list of candidates. Moreover, if we compare these baseline methods with the state-of-the-art implementations (GW@2 and SZ@2), we can see that our baselines get nearly equal or significantly higher F than those. Regarding the other baselines, BRepeated@M reports very low results both in FREESOUND and FLICKR1M datasets, and BRandom baseline stays way below all the other methods.

Number of recommended tags

Another valuable aspect to evaluate from the tag recommendation methods is the number of tags that they recommend ($|\Gamma_R|$). Table 6 shows the average $|\Gamma_R|$ for the evaluated methods using the FREESOUND and FLICKR1M datasets. We consider that methods which recommend higher number of tags and maintain overall high precision values are the most valuable for our purposes, as they provide both comprehensive and appropriate tag recommendations (i.e., relevant tags for the particular resource). In general we see that the

Table 5: Average precision P, recall R and F-measure F for tag recommendation methods using the FLICKR1M dataset, sorted by F-measure. For the sake of readability, we only show the results of baseline methods for the values of K and M that reported higher F-measure. Baseline methods are marked in italics.

	Flickr1M		
Method	**Precision**	**Recall**	**F-measure**
RankP@0.10	0.503	0.513	0.452
GW@2	*0.480*	*0.517*	*0.442*
BRankFIX@2	*0.475*	*0.511*	*0.441*
RankST	0.459	0.556	0.437
RankLR	0.384	0.597	0.414
SimP@0.20	0.462	0.422	0.394
RankKP@0.01	0.389	0.483	0.388
SimST	0.475	0.340	0.384
SimLR	0.412	0.461	0.384
BSimFIX@2	*0.417*	*0.440*	*0.382*
SZ@2	*0.384*	*0.410*	*0.353*
SimKP@0.005	0.430	0.325	0.339
BRepeated@3	*0.163*	*0.715*	*0.219*
BRandom (best)	*0.007*	*0.045*	*0.020*

best scoring methods, corresponding to the first positions of the table, recommend more tags than BRankFIX@2 and GW@2 (Table 6), and at the same time report higher (or very similar) precision values and overall F-measure (see Tables 4 and 5). If we look at the evaluation results obtained with BRankFIX@K methods when recommending more than two tags, we observe significant drops in precision ($P = 0.323$ for $K = 3$ and $P = 0.272$ for $K = 4$ in FREESOUND, and $P = 0.391$ for $K = 3$ and $P = 0.333$ for $K = 4$ in FLICKR1M). Similar precision drops are observed in GW@K ($P = 0.306$ for $K = 3$ and $P = 0.257$ for $K = 4$ in FREESOUND, and $P = 0.396$ for $K = 3$ and $P = 0.340$ for $K = 4$ in FLICKR1M). This further highlights the superiority of our proposed methods over the baselines.

It is also interesting to see that the number of recommended tags is not only driven by the selection strategy of Step 3, but also depends on the type of aggregation used in Step 2. Both in FREESOUND and FLICKR1M, we observe that when using Rank-based aggregation, highest $|\Gamma_{\mathrm{R}}|$ is obtained using the strategy of Linear regression for selecting which tags to recommend (followed by Statistical test, Percentage and Kernel percentage strategies). When using Similarity-based aggregation, the highest $|\Gamma_{\mathrm{R}}|$ is obtained with the Percentage strategy, followed by Linear regression, Statistical test and Kernel percentage strategies (Table 6). This shows that the selection strategies behave differently if the scores of Γ_{A} are ranks or similarity values. In general, Rank-based methods recommend more tags than their Similarity-based counterparts, with an average $|\Gamma_{\mathrm{R}}|$ increase of 0.38 ($p \ll 0.01$, FREESOUND) and 0.86 ($p \ll 0.01$, FLICKR1M). Given that Rank-based aggregation methods also report higher F, that reinforces the previously mentioned idea that Step 2 is better accomplished using the Rank-based strategy.

We also looked at the difference between the number of recommended tags and the

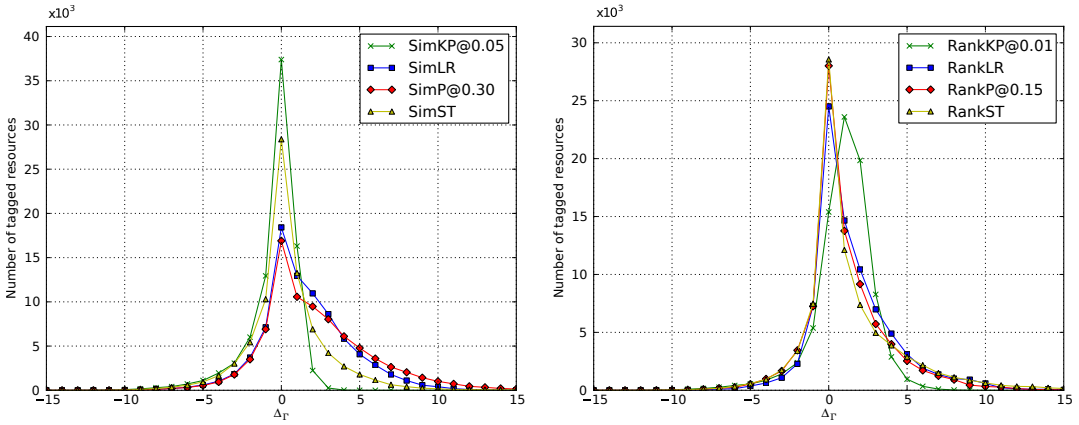

Figure 7. Histogram of the difference between the number of recommended tags and the number of deleted tags Δ_Γ for Similarity-based (a) and Rank-based (b) tag recommendation methods using FREESOUND dataset. Qualitatively similar results were obtained with FLICKR1M.

number of tags that are deleted for each resource ($\Delta_\Gamma = |\Gamma_\mathrm{R}| - |\Gamma_\mathrm{D}|$). It can be observed that most of our proposed methods report the maximum peak of the histogram of Δ_Γ at $\Delta_\Gamma = 0$ (Fig. 7). This suggests that these methods have a certain tendency to recommend as many tags as have been removed. Although it is not the goal of the tag recommendation methods to recommend the exact number of tags that have been removed (actually, this measure only makes sense under our tag prediction task-based evaluation), the results shown here are an interesting indicator that our proposed methods are able to indirectly estimate the number of deleted tags given only a set of input tags and the information embedded in the folksonomy. A plot of the average number of recommended tags as a function of the number of input tags and the number of deleted tags further supports this conclusion (Fig. 8). We can qualitatively observe how $|\Gamma_\mathrm{R}|$ grows along with $|\Gamma_\mathrm{D}|$, specially for low $|\Gamma_\mathrm{I}|$. It can also be observed that there is a tendency of $|\Gamma_\mathrm{R}|$ increasing when $|\Gamma_\mathrm{I}|$ decreases, meaning that the smaller the number of input tags, the more tags are recommended. Similar plots can be obtained with the other proposed recommendation methods, specially for RankLR and RankP (both in FREESOUND and FLICKR1M datasets).

Other relevant aspects

In order to better understand the behaviour of the proposed tag recommendation methods, we have carried out further analyses on the impact of particular aspects of the methods. To avoid very intensive computation we have only focused on the three methods that report best average F both in FREESOUND and FLICKR1M, that is to say, RankST, RankLR and RankP@α (with α being 0.15 for FREESOUND and 0.10 for FLICKR1M as shown in Table 3). The following subsections describe these experiments.

Limiting the minimum number of input tags. To assess the impact of limiting the number of input tags we now repeat the experiments including resources evaluated with less than three input tags. As we could expect, we obtain lower F scores (Table 7). In

Table 6: Average number of recommended tags $|\Gamma_R|$ for tag recommendation methods using the FREESOUND and FLICKR1M datasets (standard deviation in parentheses). Methods are displayed and sorted according to the F values of Tables 4 and 5. Baseline methods are marked in italics.

Freesound		**Flickr1M**	
Method	$\|\Gamma_\mathbf{R}\|$	**Method**	$\|\Gamma_\mathbf{R}\|$
RankP@0.15	3.03 (2.60)	RankP@0.10	2.68 (1.96)
RankST	3.36 (3.30)	*GW@2*	*2.00 (0.00)*
RankLR	3.55 (7.14)	*BRankFIX@2*	*2.00 (0.00)*
BRankFIX@2	*2.00 (0.00)*	RankST	3.96 (3.64)
RankKP@0.01	2.89 (1.29)	RankLR	4.64 (4.25)
GW@2	*2.00 (0.00)*	SimP@0.20	3.97 (1.64)
SimLR	3.42 (2.36)	RankKP@0.01	2.60 (1.47)
SimP@0.30	4.06 (3.10)	SimST	1.98 (1.70)
SimST	2.35 (2.17)	SimLR	3.15 (2.16)
SimKP@0.05	1.47 (0.70)	*BSimFIX@2*	*2.00 (0.00)*
BSimFIX@2	*2.00 (0.00)*	*SZ@2*	*2.00 (0.00)*
SZ@2	*2.00 (0.00)*	SimKP@0.05	1.35 (0.73)
BRepeated@3	*5.17 (8.17)*	*BRepeated@3*	*4.27 (3.11)*
BRandom (best)	*5.17 (8.17)*	*BRandom (best)*	*4.27 (3.11)*

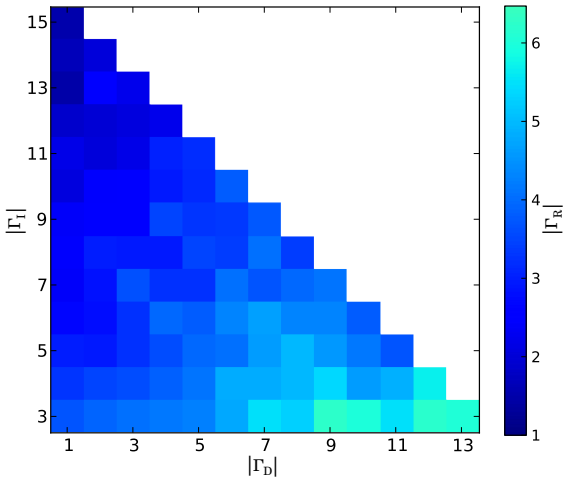

Figure 8. Average number of recommended tags $|\Gamma_R|$ as a function of the number of input tags $|\Gamma_I|$ and the number of deleted tags $|\Gamma_D|$, for method RankST, FREESOUND dataset.

Table 7: Average precision P, recall R and F-measure F for the best scoring methods in FREESOUND and FLICKR1M without filtering the number of input tags. Results are sorted in descending F score order.

Method	**Precision**	**Recall**	**F-measure**
Freesound			
RankP@0.15	0.323	0.375	0.297
RankST	0.337	0.326	0.285
RankLR	0.252	0.336	0.244
Flickr1M			
RankST	0.394	0.377	0.326
RankP@0.10	0.329	0.434	0.309
RankLR	0.244	0.352	0.243

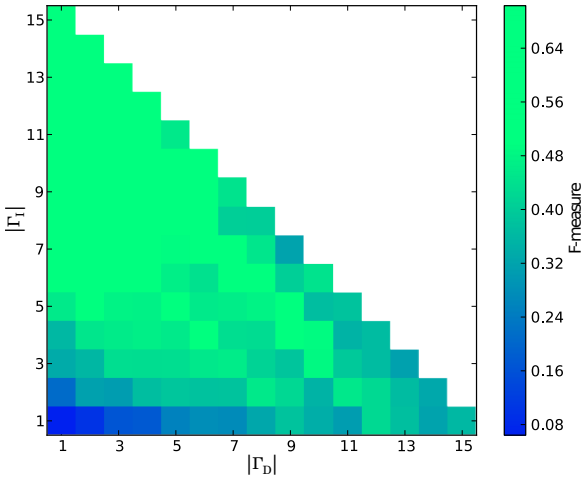

Figure 9. Average F-measure F as a function of the number of input tags $|\Gamma_R|$ and the number of deleted tags $|\Gamma_D|$ for method RankP@0.15, FREESOUND dataset. This plot includes the results of resources evaluated with less than three input tags.

average, all methods have a decrease in F of 0.154 ($p \ll 0.01$) and 0.141 ($p \ll 0.01$) for FREESOUND and FLICKR1M datasets, respectively. This confirms our initial observation that content-based methods might be more suited to recommend tags to scarcely labeled resources. In Fig. 9 we have plotted average F as a function of the number of input tags and the number of deleted tags for the RankP@0.15 method (using the FREESOUND dataset). This plot is useful to understand in which range of the number of input tags and number of deleted tags the recommendation performs better. As it can be observed, the optimum conditions for high F are found with 5 or more input tags and 6 or less deleted tags, meaning that the recommendation needs a few input tags to effectively aggregate and select candidates and not many tags to predict. Nevertheless, the fact that F is way above the random baseline of Tables 4 and 5 emphasizes that, even outside the optimum conditions, the proposed methods are still useful to some extent.

Using alternative similarity measures. As it has been explained in the evaluation methodology, all previously reported experiments have been performed using cosine similarity as the similarity measure for Step 1 of the tag recommendation scheme. In this subsection we repeat the evaluation for the best scoring methods in FREESOUND and FLICKR1M datasets, but now using Jaccard and tag co-occurrence as similarity measures (Table 8). In both datasets and for all methods, cosine similarity is the metric that obtains higher F, with an average increase of 0.009 ($p \ll 0.01$, FREESOUND) and 0.053 ($p \ll 0.01$, FLICKR1M) respect to Jaccard, and 0.086 ($p \ll 0.01$, FREESOUND) and 0.108 ($p \ll 0.01$, FLICKR1M) respect to tag co-occurrence. In the case of FREESOUND, we observe that the difference between cosine and Jaccard similarity is very small, and could be due to a marginal increase in the average number of recommended tags, thus lowering precision and getting a higher number of wrong recommendations. In FLICKR1M the increase in the average number of recommended tags is more prominent, and so it is the decrease in F for the methods using Jaccard distance. We have observed that performing the same experiment with the Similarity-based counterparts of these methods (SimP@α, SimST and SimLR) also leads to very similar results, with cosine similarity obtaining the highest F followed by Jaccard and tag co-occurrence. However, in that case the F differences among the different similarity measures tend to be slightly larger than these obtained with Rank-based methods.

Number of candidate tags per input tag (N). In order to understand the effect of N, the number of candidates per input tag (Step 1), we have performed a series of experiments with the best scoring methods for FREESOUND and FLICKR1M datasets. Similarly to the general experiments described in the evaluation strategy, we have performed 10-fold cross validations for each one of the best scoring methods, giving different values to N. To speed up computation time, we limited the number of resources of each experiment to 10,000. The rest of the parameters have remained constant (input tags in the range of $[3, 15]$, using cosine similarity, and $\alpha = 0.15$ or 0.10 for FREESOUND and FLICKR1M, respectively). The results show that most of the methods achieve a local maxima in the range of $N = [75, 150]$, and then show a very slow decaying tendency (Fig. 10). In FREESOUND, RankP@0.15 and RankST are shown to be more constant, without a noticeable decay (standard deviation of 0.005 for both RankST and RankP in the range of $N = [125, 400]$). These results suggest that after selecting a sufficient amount of N candidates for each input tag, the most relevant tags have already been selected, and increasing N does not have a big impact on the output of the recommendation as score values for the "extra" candidates are generally low. According to Fig. 10, for most of the methods, highest F-measure is obtained with $N \approx 125$, which is slightly higher than the value we used for our main experiments (N=100). However, the average F-measure increase is less than 1% and significance tests fail with $p \approx 0.10$ when comparing homologous methods with $N = 100$ and $N = 125$.

Contribution of each step. Finally, we perform several experiments to evaluate the contribution of each step of the proposed tag recommendation scheme. For the best scoring methods (RankP, RankST and RankLR) and FREESOUND and FLICKR1M datasets, we have repeated the 10-fold cross validations of the main experiments three times, replacing in each run one step of the recommendation system by a randomised version of itself.

In the first run we have replaced Step 1 by a random version that, for each input tag, selects N random candidates from the whole vocabulary of the folksonomy (using

Table 8: Average precision P, recall R, F-measure F and number of recommended tags $|\Gamma_{\mathrm{R}}|$, using different similarity measures.

Method	**P**	**R**	**F**	$\|\Gamma_{\mathbf{R}}\|$
Freesound				
Cosine similarity				
RankP@0.15	0.444	0.532	**0.437**	3.03
RankST	0.443	0.537	0.433	3.36
RankLR	0.393	0.563	0.418	**3.55**
Jaccard similarity				
RankP@0.15	0.425	0.543	**0.431**	3.28
RankST	0.421	0.552	0.423	**3.91**
RankLR	0.370	0.570	0.405	3.84
Tag Co-ocurrence				
RankP@0.15	0.339	0.483	**0.352**	3.37
RankST	0.336	0.492	0.348	3.85
RankLR	0.284	0.541	0.330	**4.65**
Flickr1M				
Cosine similarity				
RankP@0.10	0.503	0.513	**0.452**	2.68
RankST	0.459	0.556	0.437	3.96
RankLR	0.384	0.597	0.414	**4.64**
Jaccard similarity				
RankP@0.10	0.417	0.491	**0.397**	3.46
RankST	0.374	0.555	0.378	**5.97**
RankLR	0.336	0.561	0.369	5.35
Tag Co-ocurrence				
RankP@0.10	0.346	0.458	**0.337**	3.77
RankST	0.320	0.505	0.329	5.43
RankLR	0.269	0.542	0.311	**6.12**

Table 9: Average F-measures F after randomising steps 1, 2 and 3 for the best scoring tag recommendation methods in FREESOUND and FLICKR1M.

Method	**Run 1**	**Run 2**	**Run 3**	**No randomisation**
Freesound				
RankP@0.15	< 0.001	0.012	0.303	0.437
RankST	< 0.001	0.006	0.302	0.433
RankLR	< 0.001	0.007	0.302	0.418
Flickr1M				
RankP@0.10	< 0.001	0.018	0.313	0.452
RankST	< 0.001	0.010	0.313	0.437
RankLR	< 0.001	0.011	0.312	0.414

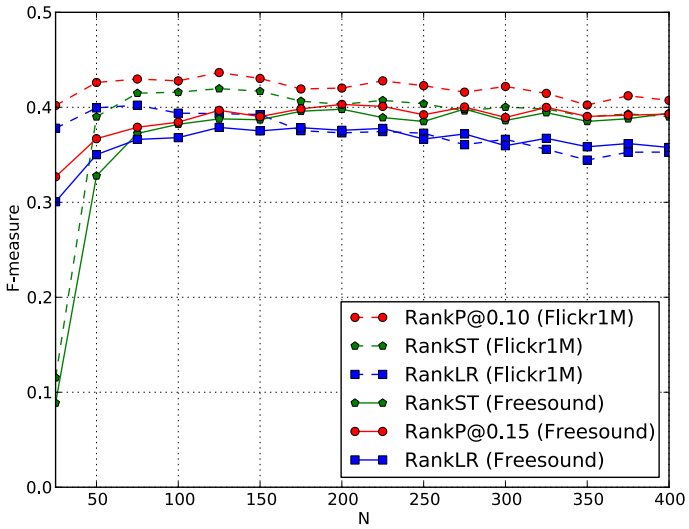

Figure 10. Average F-measure F with different values of N for the best scoring recommendation methods in FREESOUND and FLICKR1M (each experiment performed with 10,000 resources).

$N = 100$). In the second run we have maintained Step 1 as in the original setting, but have replaced Step 2 by an alternative version that, after performing a Rank-based aggregation, detaches the score values from each candidate in Γ_A, and randomly re-assigns them among the candidates. Finally, in the third run of the experiments, we have maintained Steps 1 and 2 as in the original setting but replaced the selection step by an alternative version that recommends the first K tags from Γ_A (sorted by the scores of candidates), and determines K with a random number generator with the same distribution as the number of deleted tags (which according to the main experiments can be modelled as a normal distribution with $\mu = 1.92$ and $\sigma = 1.58$ for FREESOUND, and $\mu = 2.32$ and $\sigma = 2.01$ for FLICKR1M). By applying the distribution of the number of deleted tags to the number of recommended tags, we optimize F scores as precision and recall errors are minimised when $\Delta_\Gamma \approx 0$.

Runs 1 and 2 report very low F in both datasets (Table 9). Run 3 obtains quite acceptable results, but with an average F decrease of 0.1270 ($p \ll 0.01$, FREESOUND) and 0.1214 ($p \ll 0.01$, FLICKR1M) with respect to the normally working methods (without any randomisation). Hence run 3 is still far from the optimum recommendation of normally working methods (Table 9). Given that Steps 1 and 2 are tightly coupled, failing in any of them has a very big impact on the final results. In the case of randomising Step 1, further steps can not effectively recommend tags as the original candidates are not relevant. When randomising Step 2, although candidate tags obtained in Step 1 are relevant, the aggregation can not assign meaningful scores to the candidates and thus the selection step fails in selecting which tags to recommend. Finally, when randomising Step 3, although a meaningful list of candidates can be sorted with meaningful score values, the number of tags that is recommended for each resource is selected in a completely unrelated way respect to the score distribution of the candidates, thus not considering the possible relevance of each candidate given the other candidates. Overall, this demonstrates the usefulness of each of the three proposed steps in our tag recommendation scheme.

Conclusion and discussion

In this work we have presented a general scheme for tag recommendation systems based on tag co-occurrence in folksonomies. This general scheme is composed of three steps for which we have proposed several different strategies. Step 1, Getting candidate tags, selects a number of candidate tags for every input tag based on a tag-tag similarity matrix derived from a folksonomy. Three variants of these step are given by the usage of alternative similarity measures. Step 2, Aggregating candidate tags, assigns scores to the candidates from Step 1 and merges them all in a single list of candidate tags. For that step, we have proposed two strategies which differ in the way scores are assigned. Finally, Step 3, Selecting which tags to recommend, automatically selects the candidates that will be part of the final recommendation by determining a threshold and filtering out those candidates whose score is below the threshold. For that last step we have described four strategies of different complexity levels.

From the combination of these strategies, we have proposed eight tag recommendation methods and deeply evaluated them with two real-world datasets coming from two independent collaborative tagging systems: FREESOUND and FLICKR1M. The simplicity of the described methods makes them suitable for dealing with large-scale datasets such as the ones we have used here. The most computationally expensive operation in the recommendation process is the sparse matrix multiplication performed to derive the similarity matrix, which can be done offline. Moreover, the described tag recommendation methods are easily adaptable to any other collaborative tagging system featuring a narrow folksonomy, as recommendation is solely based on tag co-occurrence information regardless of the type of resources for which tags are being recommended. Evidence for supporting this statement can be directly extracted from the qualitatively similar results achieved with the two distinct datasets employed here. We also compared our methods with simpler baselines and two state-of-the-art methods described in the literature, and analysed the effects of several parameter configurations. Our exhaustive evaluation shows that the proposed methods can effectively recommend relevant tags given a set of input tags and a folksonomy embedding

Table 10: Example of tag recommendations in FREESOUND using the RankST method. Corresponding sounds can be listened at the following url: `http://www.freesound.org/search?q=`[Sound id].

Sound id	**Input tags**	**Deleted tags**	**Recommended tags**	**F**
8780	analog, glitch, warped	lofi	noise, electronic	0.0
124021	newspaper, reading, paper, page, news	read	magazine	0.0
38006	hit, glass, oneshot	percussion	singlehit, singlebeat, single, tap, hits, house, **percussion**, place, thuds, drum, plock	0.17
54374	spring, nightingale, nature, bird	field-recording, birdsong, binaural	birds, **field-recording**, forest, **birdsong**	0.5
78282	metal, medium-loud, interaction	impact	**impact**, wood	0.67

tag co-occurrence information.

An interesting aspect of the proposed tag recommendation scheme is the step focused on automatically selecting which tags to recommend given a list of candidates. Among the four strategies we have proposed, three of them have been shown to effectively choose relevant tags for the recommendation and significantly improve the results (Percentage strategy, Statistical test strategy and Linear regression strategy). These three strategies reported similar results, though the good performance of the Statistical test and the Linear regression strategies is of special relevance as both can be considered parameter-free. We have also shown that scoring candidate tags using ranks instead of raw tag similarities statistically significantly increases the accuracy of the recommendations.

Much of the evaluation we have conducted is based on analysing the F-measure obtained after a tag prediction task. Although such systematic approach allows us to compare the different tag recommendation methods using a large number of resources, the results in terms of F-measure are probably much worse than what a user-based evaluation could have reported. To exemplify this observation, Table 10 shows a few examples of tag recommendations performed using the RankST method in the FREESOUND dataset. We have bolded the tags that are considered good recommendations under our evaluation framework. Notice however that many of the recommended tags which are not bolded could also be judged as meaningful recommendations if we actually listen to the sounds (Table 10). Moreover, our systematic evaluation does not take into account other aspects of the recommended tags such as their semantic context or their informational value in the folksonomy. We think that future research should consider these kind of aspects, and tag recommendation systems should include domain-specific knowledge to provide more meaningful recommendations.

Constructing better tag recommendation systems yielding more useful tag descriptions of online resources would allow improved organisation, browsing and reuse of online content. But not only that. In the literature on social and semantic web there have been many discussions regarding the relevance and value of folksonomies compared to ontologies and vice versa (Shirky, 2005; Mika, 2007; Gruber, 2007; Al-Khalifa & Davis, 2007). In some cases, both concepts appear as opposed approaches for bottom-up (folksonomies) or top-down (ontologies) knowledge sharing. However, many authors coincide in that both approaches can coexist and reciprocally benefit from each other. In this direction, many studies have been performed around the idea of extracting structured knowledge from folksonomies (Limpens et al., 2009). We too believe that folksonomies are a powerful tool from which relevant structured knowledge can be gathered, and that they have to play a very important role in the semantic web. By using approaches such as the one we have presented here, more coherent and less noisy folksonomies can emerge and we can help leveraging their value as reliable sources for knowledge-mining and ontology-induction.

References

Akkermans, V., Font, F., Funollet, J., Jong, B. D., Roma, G., Togias, S., & Serra, X. (2011). Freesound 2: An improved platform for sharing audio clips.. Late-breaking demo abstract

of the International Society for Music Information Retrieval Conference. Retrieved from http://mtg.upf.edu/node/2376

Al-Khalifa, H. S., & Davis, H. C. (2007). Exploring the value of folksonomies for creating semantic metadata. *International Journal on Semantic Web and Information Systems*, *3*(1), 13-39.

Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D., & Jordan, M. (2003). Matching words and pictures. *Journal of Machine Learning Research*, *3*, 1107-1135.

Barrington, L., Chan, A., Turnbull, D., & Lanckriet, G. (2007). Audio information retrieval using semantic similarity. In *Proceedings of the ieee international conference on acoustics, speech and signal processing* (Vol. 2, p. 725-728).

Bischoff, K., Firan, C., Nejdl, W., & Paiu, R. (2008). Can all tags be used for search? In *Proceedings of the 17th acm conference on information and knowledge management* (p. 193-202).

Cantador, I., Konstas, I., & Jose, J. (2011). Categorising social tags to improve folksonomy-based recommendations. *Journal of Web Semantics*, *9*(1), 1-15.

Cao, H., Xie, M., Xue, L., & Liu, C. (2009). Social tag prediction base 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* (p. 35-48).

Farooq, U., Kannampallil, T. G., Song, Y., Ganoe, C. H., Carroll, J. M., & Giles, L. (2007). Evaluating tagging behavior in social bookmarking systems: Metrics and design heuristics. In *Proceedings of the acm international conference on supporting group work* (p. 351-360).

Flickr. (2012). In *Wikipedia.* Retrieved 2/10/2012, from http://en.wikipedia.org/wiki/Flickr

Garg, N., & Weber, I. (2008). Personalized, interactive tag recommendation for flickr. In *Proceedings of the 2nd acm conference on recommender systems* (p. 67-74).

Golder, S. A., & Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science*, *32*(2), 198-208.

Gruber, T. (2007). Ontology of folksonomy: A mash-up of apples and oranges. *International Journal on Semantic Web and Information Systems*, *3*(1), 1-11.

Halpin, H., Robu, V., & Shepard, H. (2006). The dynamics and semantics of collaborative tagging. In *Proceedings of the 1st semantic authoring and annotation workshop* (p. 1-21).

Hotho, A., Jäschke, R., Schmitz, C., & Stumme, G. (2006). Information retrieval in folksonomies: Search and ranking. In *Proceedings of the 3rd european semantic web conference* (p. 411-426).

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* (p. 497-506).

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* (p. 506-514).

Kruskal, W. H., & Wallis, W. A. (1997). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*(260), 583-621.

Li, J., & Wang, J. Z. (2006). Real-time computerized annotation of picture. In *Proceedings of the 14th acm multimedia conference* (p. 911-920).

Limpens, F., Gandon, F. L., & Buffa, M. (2009). *Linking folksonomies and ontologies for supporting knowledge sharing: a state of the art.* Retrieved from http://isicil.inria.fr/v2/res/docs/livrables/ISICIL-ANR-EA01-Folksonomies Ontologies-0906.pdf

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* (p. 84-95).

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* (p. 7-15).

Markines, B., Cattuto, C., Menczer, F., Benz, D., Hotho, A., & Stumme, G. (2009). Evaluating similarity measures for emergent semantics of social tagging. In *Proceedings of 18th international world wide web conference* (p. 641-650).

Marlow, C., Naaman, M., Boyd, D., & Davis, M. (2006). Ht06, tagging paper, taxonomy, flickr, academic article, toread. In *Proceedings of the 17th acm conference on hypertext and hypermedia* (p. 31-41).

Martínez, E., Celma, O., Sordo, M., Jong, B. D., & Serra, X. (2009). Extending the folksonomies of freesound.org using content-based audio analysis. In *Proceedings of the sound and music computing conference* (p. 23-25).

Meo, P. D., 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. *Information Systems Journal*, *34*(6), 511-535.

Mika, P. (2007). Ontologies are us: A unified model of social networks and semantics. *Journal of Web Semantics*, *5*(1), 5-15.

Papadopoulos, S., Kompatsiaris, Y., & Vakali, A. (2010). A graph-based clustering scheme for identifying related tags in folksonomies. In *Proceedings of the 12th international conference on data warehousing and knowledge discovery* (p. 65-76).

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* (p. 235-242).

Salzberg, S. L. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach. *Journal of Data Mining and Knowledge Discovery*, *1*(3), 317-328.

Scholz, F. W., & Stephens, M. A. (1987). K-sample anderson-darling tests. *Journal of the American Statistical Association*, *82*(399), 918-924.

Scott, D. W. (2008). Multivariate density estimation: Theory, practice, and visualization. In (p. 125-193). John Wiley & Sons.

Shirky, C. (2005). *Ontology is overrated: Categories, links, and tags.* Retrieved 24/10/2012, from `http://www.shirky.com/writings/ontolog_overrated.html`

Sigurbjörnsson, B., & Zwol, R. V. (2008). Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th international conference on world wide web* (p. 327-336).

Silvermann, B. W. (1986). Density estimation for statistics and data analysis. London: Chapman & Hall/CRC.

Sordo, M. (2012). *Semantic annotation of music collections: A computational approach.* Unpublished doctoral dissertation, Universitat Pompeu Fabra, Barcelona.

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*(2), 467-476.

Wal, T. V. (2005). *Explaining and showing broad and narrow folksonomies.* Retrieved 26/09/2012, from `http://www.vanderwal.net/random/entrysel.php?blog =1635`

Wu, L., Yang, L., Yu, N., & Hua, X.-S. (2009). Learning to tag. In *Proceedings of the 18th international conference on world wide web* (p. 361-370).