Universitat Pompeu Fabra
Barcelona

Tag Recommendation using Folksonomy Information for Online Sound Sharing Platforms

Frederic Font Corbera

TESI DOCTORAL UPF / 2015

Directors de la tesi:

Dr. Xavier Serra Casals
Dept. of Information and Communication Technologies
Universitat Pompeu Fabra, Barcelona, Spain

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

Dissertation submitted to the Deptartment of Information and Communication Technologies of Universitat Pompeu Fabra in partial fulfillment of the requirements for the degree of

DOCTOR PER LA UNIVERSITAT POMPEU FABRA

Music Technology Group (`http://mtg.upf.edu`), Dept. of Information and Communication Technologies (`http://www.upf.edu/dtic`), Universitat Pompeu Fabra (`http://www.upf.edu`), Barcelona, Spain.

Acknowledgements

I can say now, after four and a half years working on this thesis, that so far these have been the greatest and most challenging years of my life. I learned a lot of things during these years, and surely not only about *tags*. There are many people I want to thank for having contributed, in one way or another, to make this happen. First of all, I would like to thank Xavier Serra, not only for giving me the opportunity to join the MTG and supervising this thesis, but also for having lead the MTG for more than 20 years and always being enthusiastic about new projects and ideas. I remember, when I was about to apply for the grant that then supported my research, Xavier told me that pursuing a PhD is a "way of life". He was right, and has helped me since then. However, there is someone else to whom I feel much obliged for helping me out. After I did my first presentation on what my thesis was going to be about, Joan Serrà told me that he was interested and, if I wanted, he could help me. I was not aware at the time of how important that collaboration would become, nor about how much I would learn from him. Fortunately I knew enough to say *yes*, and so Joan became co-supervisor of the thesis. Since then, his help at all stages has been invaluable.

A very important element of this thesis has been Freesound. Working with Freesound has been a great motivation throughout the thesis, and it has turned me into a developer that now reads programming books and Python blogs. All because I have been lucky enough to be part of the Freesound team, and to work with people like Gerard Roma, Alastair Porter, Bram de Jong, and former Freesound team members Vincent Akkermans, Stelios Togias and Jordi Funollet. To all of you, I sincerely thank you for teaching me so many *geeky* things, research and programming philosophy. Also related to Freesound, I want to particularly thank the Freesound moderators and all Freesound users that participated in my online experiments and that contribute everyday to make Freesound such an amazing site.

There is many other people at the MTG without whom this journey would not have been half as enjoyable. My lunch mates from the "Lunch time conversations (official thread)" Skype group: Panos Papiotis, Sergio Oramas, Sebastian Mealla, Oriol Romaní, Giuseppe Bandeira, Dara Dabiri, Álvaro Sarasúa, Juanjo Bosch, Martí Umbert, Carles F. Julià, and everyone else with whom I shared thoughts on the thesis or about any other professional or personal aspects: Sankalp Gulati, Sertan Şentürk, Mohamed Sordo, Gopala Krishna Koduri, Dmitry Bogdanov, Rafael Caro, Agustín Martorell , Nadine Kroher, Ajay Srinivasamurthy, Justin Salamon, and those that I'm missing! From the MTG, I would also particularly like to thank Perfecto Herrera for his

input when designing user experiments, and Alba Rosado for making me feel I could do much more than research at the MTG. I also want to thank Tamsin Porter for proofreading this thesis.

Last but not least, I want to thank all my friends and family and, with a particular emphasis, I would like to thank Anna for always being there, accompanying me throughout the whole process, and letting me accompany her on her own. Thank you!

This thesis has been carried out at the Music Technology Group of Universitat Pompeu Fabra (UPF) in Barcelona, Spain, from October 2010 to February 2014 and from May 2014 to March 2015, and at the Centre for Digital Music of Queen Mary University of London (QMUL), United Kingdom, from March 2014 to April 2014. This work has been supported by the Spanish Ministry of Science and Innovation (BES-2010-037309 FPI grant and TIN-2009-14247-C02-01 DRIMS project), and by the European Research Council (FP7-2007-2013 / ERC grant agreement 267583). The research stay at QMUL has been also funded by the Spanish Ministry of Science and Innovation (EEBB-I-14-08838).

Abstract

Online sharing platforms host a vast amount of multimedia content generated by their own users. Such content is typically not uniformly annotated and can not be straightforwardly indexed. Therefore, making it accessible to other users poses a real challenge which is not specific of online sharing platforms. In general, content annotation is a common problem in all kinds of information systems. In this thesis, we focus on this problem and propose methods for helping users to annotate the resources they create in a more comprehensive and uniform way. Specifically, we work with tagging systems and propose methods for recommending tags to the content creators during the annotation process. To this end, we exploit information gathered from previous resource annotations in the same sharing platform, the so called *folksonomy*. Tag recommendation is evaluated using several methodologies, with and without the intervention of users, and in the context of large-scale tagging systems. We focus on the case of tag recommendation for sound sharing platforms. Besides studying the performance of several methods in this scenario, we analyse the impact of one of our proposed methods on the tagging system of a real-world and large-scale sound sharing site. As an outcome of this thesis, one of the proposed tag recommendation methods is now being daily used by hundreds of users in this sound sharing site. In addition, we explore a new perspective for tag recommendation which, besides taking advantage of information from the folksonomy, employs a sound-specific ontology to guide users during the annotation process. Overall, this thesis contributes to the advancement of the state of the art in tagging systems and folksonomy-based tag recommendation, and explores interesting directions for future research. Even though our research is motivated by the particular challenges of sound sharing platforms and mainly carried out in that context, we believe our methodologies can be easily generalised and thus be of use to other information sharing platforms.

Resum

Les plataformes d'intercanvi de recursos multimèdia a Internet contenen grans quantitats de contingut creat pels seus usuaris. Habitualment, aquest contingut no està ben anotat, i això fa que la seva indexació no sigui una tasca fàcil. Aconseguir que aquest contingut sigui accessible pels altres usuaris suposa un repte important, el qual no és només específic d'aquest tipus de plataformes. En general, l'anotació de contingut és un problema comú en molts tipus de sistemes d'informació. En aquesta tesi, ens focalitzem en aquest problema i proposem mètodes per ajudar els usuaris a anotar, d'una manera més completa i uniforme, el contingut creat per ells mateixos. Concretament, treballem amb sistemes d'etiquetatge – *tagging* – i proposem mètodes per recomanar etiquetes – *tags* – durant el procés d'anotació del contingut. Per aconseguir això, analitzem la manera com els altres continguts de la plataforma d'intercanvi han estat etiquetats prèviament. Aquesta informació s'anomena *folksonomia*. Avaluem la tasca de recomanar tags utilitzant diverses metodologies, amb o sense la participació d'usuaris, i en el context de sistemes de tagging a gran escala. Particularment, ens focalitzem en el cas de la recomanació de tags en plataformes d'intercanvi de sons i, a part de testar el funcionament de diferents mètodes en aquest escenari, també analitzem l'impacte d'un d'aquests mètodes en el sistema de tagging d'una plataforma d'intercanvi de sons real. De fet, de resultes d'aquesta tesi, centenars d'usuaris fan servir diàriament un dels sistemes proposats de recomanació de tags en aquesta plataforma d'intercanvi. A més a més, també explorem un nou enfocament per als sistemes de recomanació de tags que, a part de nodrir-se de la informació de la folksonomia, incorpora una ontologia amb informació sobre l'àmbit del so que serveix per guiar els usuaris durant el procés d'anotació de contingut. En general, aquesta tesi contribueix a l'avenç de l'estat de l'art dels sistemes de tagging i de recomanació de tags basats en folksonomies, i explora direccions interessants per continuar investigant. Tot i que la nostra recerca està motivada pels reptes particulars que proposen les plataformes d'intercanvi de sons i està avaluada principalment en aquest context, creiem que les metodologies que proposem poden ser generalitzades fàcilment i utilitzades en altres plataformes d'intercanvi.

Contents

List of figures

List of tables

List of mathematical symbols

General

Example	Symbol type	Description
a, b, γ	Lowercase letters	Indices, variables, vector, set and matrix elements.
A, B, Γ	Uppercase letters	Constants, functions and evaluation metrics.
$\mathbf{A}, \mathbf{B}, \mathbf{C}$	Bold uppercase letters	Vectors and sets.
$\mathcal{A}, \mathcal{B}, \mathcal{C}$	Calligraphy letters	Graphs, matrices and other complex data structures.

Specific

Symbol	Description
$\mathbf{A}$	Set of annotation sessions
a	Element of $\mathbf{A}$ (a particular annotation session)
$\mathbf{C}$	Set of audio classes
$\mathcal{D}$	Association matrix
d	Element of $\mathcal{D}$
$\mathbf{E}$	Vector of tag applications (edges of the folksonomy hypergraph)
F	F-measure evaluation metric
$\mathcal{F}$	Folksonomy
I	(In)coherence in annotations evaluation metric
M	Percentage of misspelled tag applications evaluation metric
$\mathcal{O}$	Ontology
P	Precision evaluation metric
p	p-value in statistical tests
Q	Subjective annotation quality evaluation metric
$\mathbf{Q}$	Union of all qualitative judgements of sound annotations
q	Qualitative judgement for the annotation of a sound
$\mathcal{R}$	Sound-sound graph
R	Recall evaluation metric
$\mathbf{R}$	Set of resources (typically of sounds)
r	A particular resource (typically a sound)
$\mathcal{S}$	Tag-tag similarity matrix

Symbol	Description
s	Element of $\mathcal{S}$
$\mathbf{T}$	Set of tags
$\mathbf{T}_{\mathrm{A}}$	Set of aggregated candidate tags
$\mathbf{T}_{\mathrm{C}}$	Set of candidate tags
$\mathbf{T}_{\mathrm{D}}$	Set of deleted tags
$\mathbf{T}_{\mathrm{I}}$	Set of input tags
$\mathbf{T}_{\mathrm{R}}$	Set of recommended tags
$\mathbf{T}^r$	Set of tags assigned to a resource r (tagline of the resource)
$\mathbf{T}_{\mathrm{T}}$	Set of attribute-tags
$\mathbf{T}_{\mathrm{Z}}$	Set of tags populated under a tag category
t	A particular tag
$\mathcal{TR}$	Bipartite graph relating tags and resources
$\mathcal{U}$	User-user graph
$\mathbf{U}$	Set of users
u	A particular user
W_I	Analysis window of interest
$\mathbf{W}$	Vector of reference analysis windows
w	Edge weight for user-user and sound-sound graphs
$\mathbf{Z}$	Set of tag categories
$\mathbf{Z}_{\mathrm{R}}$	Set of recommended tag categories
z	Element of $\mathbf{Z}$ (a particular tag category)
α	Percentage parameter of Percentage Strategy
β	Percentage parameter of Kernel Percentage Strategy
Γ	Average tagline length evaluation metric
ε	Score threshold for candidate tags
Θ	Average percentage of new tags evaluation metric
θ	Number of candidate tags per input tag
κ	Fixed number of recommended tags
Λ	Annotation comprehensiveness evaluation metric
λ	Duration of an annotation session
Ψ_{u}	User vocabulary sharing evaluation metric
Ψ_{r}	Sound vocabulary sharing evaluation metric
Φ_{e}	Average tag application time evaluation metric
Φ_{r}	Average time per sound evaluation metric
ϱ	Number of repeated tags in Repeated aggregation and selection strategy
Υ	Average user vocabulary size evaluation metric
υ	Tag frequency of occurrence
ϕ	Score of a candidate tag
Π	Average percentage of attribute-tags evaluation metric
Ω	Average number of correctly predicted tags evaluation metric
ω	Tag frequency threhsold

CHAPTER 1

Introduction

1.1 Motivation

Information sharing is considered by some authors as being an "attribute of humanity itself" (Dunbar, 1998; Rafaeli & Raban, 2005). In the last two decades, the web has revolutionized the way in which information is shared among human beings. The so called *social media* revolution (Smith, 2009) has brought sharing to a whole new level. Nowadays, all kinds of information such as books, articles, opinions, videos, pictures and audio tracks are hosted in online sharing platforms accessed by millions of users worldwide. Besides the fact that the amount of information that is accessible through the internet is huge, far bigger than what was available through traditional channels (i.e., libraries, television, radio, shops, etc.), there is something very particular in the online sharing paradigm: much of the content that is shared online is generated by the users of these sharing platforms (Kietzmann et al., 2011), the so called *user generated content* (Kaplan & Haenlein, 2010).

Such user generated content potentially represents an incredibly valuable resource that can serve several purposes, ranging from business and research applications to artistic creation and the preservation of cultural heritage (Krumm et al., 2008). Nevertheless, the value of user generated content is significantly dimmed by the ways in which such content can be accessed and reused. As the amount of content grows, so does the difficulty of browsing and locating what one needs, and so do the challenges that search engines have to face.

For the content to be accessible, it needs to be properly indexed. However, the quantity and variety of user generated content turns proper indexing into a very difficult task. This is particularly true for multimedia resources like video, pictures and audio – typically the most popular in the user generated content world – which, as opposed to other kinds of media, do not have a direct textual representation (Bischoff et al., 2008). Moreover, as user generated content does not normally follow a traditional editorial process before publication, no standard metadata is generated that can help the indexing process. At the

same time, the amount of content generated is simply too much to be curated in scalable ways by groups of experts (Mathes, 2004).

Online sharing platforms typically delegate the responsibility of describing or annotating[1] their content to the users, that is to say, the *authors* or *contributors*. Thus, content description is done in a distributed fashion. The nature of content annotations may vary depending on each particular sharing platform, and is highly dependent on the description mechanism used in every particular site. Description mechanisms that look for the most uniform annotations can use forms with a number of predefined fields with fixed responses that users need to choose from when uploading a resource. For example, users might be asked to select a music genre from a particular list when uploading an audio track to an audio sharing platform. However, these mechanisms lack flexibility when new resources are uploaded, as their characteristics can be unexpected and not contemplated in the description form (Mathes, 2004; Shirky, 2005; Halpin et al., 2006; Macgregor & Mcculloch, 2006). Other description mechanisms provide more flexibility by not limiting form fields to a specific set of responses. In that case, annotations typically consist of a textual description plus a list of labels assigned by the users which are not restricted to a particular vocabulary. These labels are commonly known as *tags*, and act as keywords that describe the content. In both cases, users may apply their own ideas and rationale to decide how to describe the content, working at different levels of abstraction, and understanding the goal of the annotation process in different ways (Golder & Huberman, 2006). If we were to ask two different users to independently annotate a single online resource, we would most probably find little overlap in their responses. This illustrates the so called *vocabulary problem* (Furnas et al., 1987), which clearly shows up in today's online sharing platforms (Marlow et al., 2006).

Consequently, the organisation, browsing and searching capabilities of online sharing platforms is rather limited, particularly of those focused on multimedia sharing. For example, in YouTube[2], SoundCloud[3] or Flickr[4] (well known sites for sharing video, music and photos, respectively), the main way of accessing content is by introducing some textual query terms to be matched against resources' metadata (i.e., filenames, textual descriptions, tags, etc.). In some cases, search results can be filtered using basic file properties (i.e., duration, size, data format, etc.) and simple metadata (i.e., upload date, license, geolocation, etc.), or augmented through automatic resource recommendation systems and information from social connections (i.e., friends, followed users, etc.). YouTube and SoundCloud also offer the option to filter content using a

[1] In this thesis, the terms "annotation" and "description" can both be used interchangeably, and may refer to all kinds of metadata that can accompany an online resource.

[2] `http://www.youtube.com`. All URLs in this thesis were last accessed on 11 March 2015.

[3] `http://www.soundcloud.com`.

[4] `http://www.flickr.com`.

number of predefined categories for kinds of videos and music genres, respectively. These categories can be filled in by users when describing their content during the upload process. However, no comprehensive faceted or hierarchical browsing functionalities are possible, and no advanced search filters can be defined that can operate at a higher semantic level. For example, it would be desirable to reliably filter resources according to objects appearing in pictures or videos, or musical instruments in audio tracks.

In order to mitigate the annotation problems, substantial research has been conducted to derive computational methods for automatically annotating multimedia resources. These methods are based on the analysis of resources's content, that is to say, pixel values in pictures and video, and the audio waveform in sound and music. For example, methods have been proposed for recognizing semantic concepts in pictures (Li & Wang, 2008), for identifying human actions in videos (Poppe, 2010), for automatically classifying audio tracks in musical genres (Scaringella et al., 2006) or for the identification of environmental sounds (Chachada & Kuo, 2013). Some of these methods can achieve reasonably high accuracies and potentially allow for a uniform annotation of resources without the need of user intervention. However, these are generally still far from satisfactory in real-world scenarios (Wang et al., 2012).

It is the rationale of this thesis that if we concentrate on acquiring better content annotations from the authors themselves, we can obtain more accurate descriptions at higher semantic levels that could hardly be obtained otherwise using current content-based approaches. Hence, the idea of focusing on the acquisition of better annotations from users that upload content to online sharing sites is the starting point of our work. In particular, in this thesis we propose methods aimed at the improvement of tag annotations of user generated content.

1.2 Tagging systems and folksonomies

Using tags as keywords for annotating resources has become standard practice in online sharing sites. Tags, as a common form of user provided metadata, were first introduced in the bookmark sharing site Delicious[5] (Gupta et al., 2010; Wikipedia, 2014b), and have been adopted by many other sites. To name a few examples, sites that incorporate tagging systems include Flickr (photo sharing), YouTube and Vimeo[6] (video sharing), CiteULike[7] and Mendeley[8] (for sharing scholarly references), SoundCloud, Last.fm[9] and Freesound[10] (au-

[5] http://www.delicious.com.
[6] http://www.vimeo.com.
[7] http://www.citeulike.org.
[8] http://www.mendeley.com.
[9] http://www.last.fm.
[10] http://www.freesound.org.

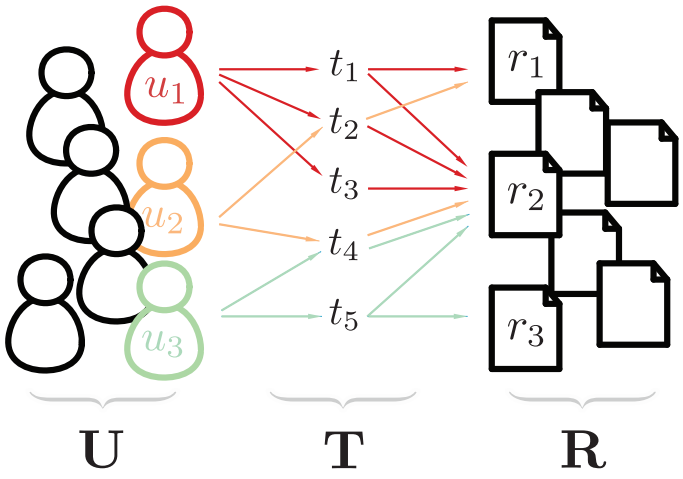


Figure 1.1: Conceptual diagram of tagging systems. **U**, **T** and **R** correspond to a set of users, tags and resources, respectively, while u_i, t_i and r_i correspond to individual users, tags and resources.

dio and music sharing), StackExchange[11] (question and answer sites), and Blogger[12] and Wordpress[13] (blogging sites).

Tags are free-form textual labels that convey some semantically meaningful information about the content resource to which they are assigned. Users apply these labels to content resources, hence *tagging* is the action that a user performs when assigning a tag to a resource[14]. In Fig. 1.1, we show a conceptual diagram for tagging systems. In it, we can observe that every tag $t \in \mathbf{T}$ can be related with a number of users $u \in \mathbf{U}$ and resources $r \in \mathbf{R}$. Every unique tag-user-resource ternary relation (e.g., the path that relates t_1, u_1 and r_1 in Fig. 1.1), is known as a *tag application* (Sen et al., 2006). In this thesis, we will refer to the union of all tags related to a particular resource as the resource's *tagline*. In the example of Fig. 1.1, the tagline of r_1 corresponds to $\{t_1, t_2\}$, whereas the tagline of r_2 corresponds to $\{t_1, t_2, t_3, t_4, t_5\}$.

The aggregate of all tag applications which relate the tags, users and resources of a sharing site, is commonly known as a *folksonomy* (Vander Wal, 2007). This term has been extensively used in the tagging systems literature, not only to designate the set of tags, users, resources and ternary relations of a sharing site, but also to refer to the implicit information and knowledge embedded in this space. However, its original meaning, as introduced by Vander Wal (2007), is somewhat more restrictive and particularly refers to those tagging systems

[11] `http://www.stackexchange.com`.
[12] `http://www.blogger.com`.
[13] `http://www.wordpress.org`.
[14] In this thesis we also use the terms "tagger" or "annotator" to refer to a user that assigns tags to a resource.

in which tag applications are made by the consumers of online resources, not necessarily the authors of the content (see below). In folksonomies, unlike more rigid structures such as taxonomies or ontologies, the semantic terms represented by tags are organized without hierarchy, and are a close representation of the vocabulary of the users in an online sharing platform (Gupta et al., 2010). In fact, the set of tags of a folksonomy is commonly known as the *vocabulary* of the folksnonomy or of a tagging system.

There are several types of tagging systems which differ in their design and purpose. Generally, the types of content resources that are shared bring some important implications regarding the design of tagging systems. On the one hand, in those sites in which users share *references* to already existing resources (e.g., links to web pages), tags are typically added by the users that consume these resources to allow further easy retrieval. As a result, a single resource can be tagged by many users, and a single tag can be assigned more than once to the same resource. The folksonomy resulting in tagging systems of this type is called a *broad folksonomy* (Vander Wal, 2005), and is the one that is more closely related to the original definition of folksonomy. Delicious, Last.fm and CiteULike are examples of online sharing sites featuring broad folksonomies. On the other hand, in sites where user generated content is shared, tags are typically assigned by the authors of the resources (i.e., the users that upload the resources) so that these are accessible to other users (Cattuto, 2006). In that case, the resulting folksonomy is called a *narrow folksonomy* (Vander Wal, 2005), and tags are only assigned once to particular resources. Examples of tagging systems with narrow folksonomies include multimedia sharing sites like Flickr, YouTube, Soundcloud and Freesound.

In the tagging systems literature, it is very common to use the terms *social tagging* and *collaborative tagging* to refer to tagging systems in general. Although the terms are normally treated as being exchangeable, their original meanings have some implications. The term collaborative tagging, first introduced by Golder & Huberman (2006), specially refers to these tagging systems in which resources can be tagged by any user (not only the authors), thus resulting in broad folksonomies. The term social tagging, introduced by Marlow et al. (2006), refers to sharing sites in which tags are particularly exposed to consumers and shared among contributors, and not only used for the self organisation of contributors' resources. In this thesis, we employ the more generic term "tagging systems". The annotation of resources using tagging systems allows online sharing platforms to provide a number of functionalities for indexing, searching and browsing their content. For example, using a *tagcloud* (Fig. 1.2), users can have an idea of the most popular tags used in a tagging system and navigate among resources by filtering their tags. In a way, the folksonomy can be used as a "semantic map" to navigate the contents of a sharing site (Cattuto, 2006).

Despite the popularity of tagging systems and their successful implementa-

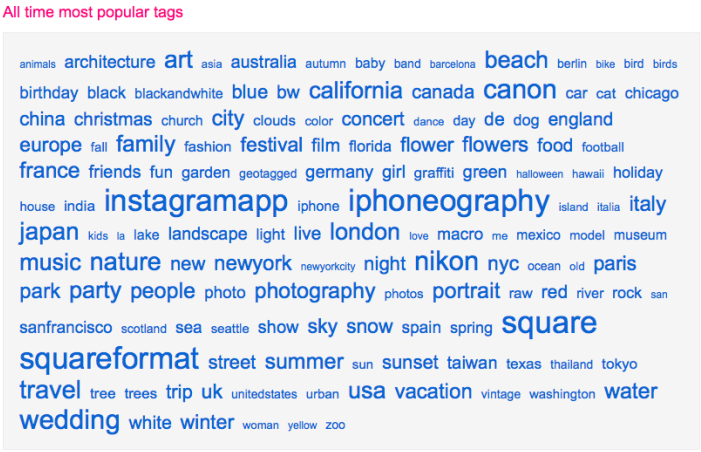


Figure 1.2: Flickr tagcloud (as retrieved on 24 July 2014).

tion in many online sharing sites, there are a number of well-known problems which limit the possibilities of these functionalities (Guy & Tonkin, 2006). These problems range from the use of different tags to refer to a single concept (synonymy) and the ambiguity in the meaning of certain tags (polysemy), to tag scarcity and the appearance of typographical errors (Golder & Huberman, 2006; Halpin et al., 2006). Furthermore, the quality of the indexing, searching and browsing functionalities enabled by tagging systems strongly relies on the coherence and comprehensiveness of the tags assigned to the resources. This does not only apply at the scope of a particular tagline of a resource, but also at the scope of the whole folksonomy (Spiteri, 2007). In other words, it is not only important that individual resources are properly tagged, but also that there is a coherence across the descriptions of all resources in the folksonomy. For that reason, it has been often discussed whether the folksonomy of a tagging system, after a certain time of being in use, reaches a point of implicit consensus where the vocabulary converges to a certain set of tags and tagging conventions that are widely adopted by all users of the system (Halpin et al., 2006; Sen et al., 2006; Sood et al., 2007; Robu et al., 2009; Wagner et al., 2014). According to these authors, the point of consensus may be reached because of imitation patterns and users' shared cultural knowledge. Reaching that point of consensus is desirable to improve the browsing and searching experience of a site (Guy & Tonkin, 2006).

For addressing some of the aforementioned problems of folksonomies, the literature of tagging systems has often proposed the implementation of tag recommendation methods to aid users in the tagging process (Golder & Huberman, 2006; Halpin et al., 2006; Marlow et al., 2006). By using such methods, user annotations are expected to be more uniform and comprehensive. The study of tag recommendation systems is the main topic of this thesis.

1.3 Tag recommendation

Tag recommendation systems are used to suggest potentially relevant tags to users when they are annotating online resources. In this way, tag recommendation assists users during the annotation process, and can have a considerable impact on the resulting annotations and folksonomies. Depending on the type of information that a tag recommendation system employs when suggesting tags, systems can be generally categorized into three main groups (Wang et al., 2012):

- The first group corresponds to the systems that, given a user u annotating a resource r, analyse the content of the resource r and automatically extract a number of features that can be related to a set of tags $\mathbf{T}_{\mathrm{R}}$ which are finally recommended. For example, given an image file, an automatic system could be used to try to automatically recognize an object appearing in the image and then the name of that object could be suggested to the user as a tag. We refer to these kind of systems as content-based tag recommendation systems.

- The second group corresponds to the systems that, to generate $\mathbf{T}_{\mathrm{R}}$, take into account the folksonomy of the tagging system and a set of tags $\mathbf{T}_{\mathrm{I}}$ that have already been assigned to the resource r. In that case, the system could recommend tags that are popular or that have been frequently used alongside the tags in $\mathbf{T}_{\mathrm{I}}$. These systems are known as folksonomy-based tag recommendation systems.

- The third group includes recommendation systems which rely on other types of metadata and contextual information to generate the recommendations. Such systems would, for example, use the title of a resource or any associated geolocation metadata to query an external service like a search engine and extract some keywords from the results to generate $\mathbf{T}_{\mathrm{R}}$. We refer to these kind of systems as context-based tag recommendation systems.

In some cases, tag recommendation systems incorporate sources of information belonging to more than one group. For example, a tag recommendation system can combine content analysis and folksonomy information to generate a list of tag suggestions. The diagram in Fig. 1.3 illustrates the concept of tag recommendation and the different kinds of information sources that can be used in recommendation algorithms.

An important advantage of content-based tag recommendation systems is that they can generate tag recommendations for resources that do not have any kind of metadata. Conversely, folksonomy-based tag recommendation requires the existence of at least one tag assigned to a resource r in order to generate recommendations tailored to that particular resource r, and context-based

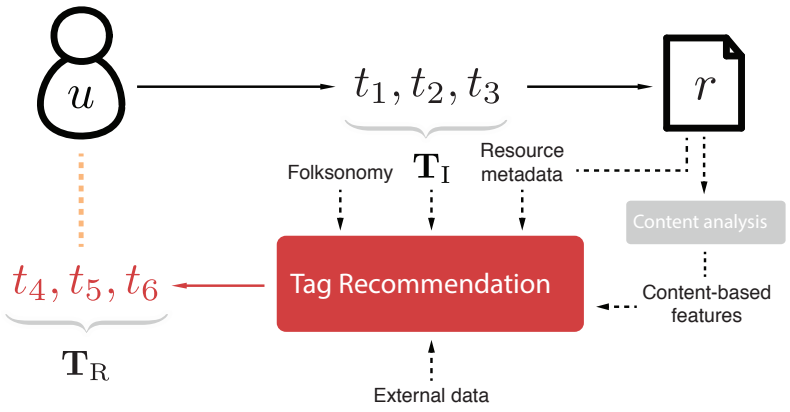


Figure 1.3: Conceptual diagram of tag recommendation systems. $\mathbf{T}_\mathrm{R}$ corresponds to a set of recommended tags, while $\mathbf{T}_\mathrm{I}$ corresponds to a set of input tags. Similarly to Fig. 1.1, u, t and r correspond to individual users, tags and resources, respectively.

tag recommendation systems also require at least a partial annotation of the resources in order to provide tag suggestions. Nevertheless, folksonomy-based and context-based tag recommendation systems have the advantage of not requiring any specific processing of the content of resources being annotated, thus typically being less computationally expensive and more easily generalisable to other types of resources. Among the different types of tag recommendation systems described above, this thesis is focused on those based on the analysis of the folksonomy of a tagging system.

Once a tag recommendation system generates a set of recommended tags $\mathbf{T}_\mathrm{R}$ for a resource r, these are normally presented as a suggestion to the user that is annotating r. Users can then decide whether or not to add any of the tags in $\mathbf{T}_\mathrm{R}$ as an annotation of r. Hence, the user acts as a judge for the quality and appropriateness of the recommendations. Intuitively, a good tag recommendation system suggests tags that users tend to consider as appropriate and therefore are added to the resource tagline. However, besides suggesting relevant tags, tag recommendation systems can also bring other benefits to the annotation process. Even when suggested tags are not considered relevant, these may inspire or convey the sorts of information that users should annotate (Ames & Naaman, 2007). For example, when annotating a music track, a tag recommendation system could recommend a tag that describes a music genre like `pop`. This tag might not be relevant for the particular music track, but it might remind the user to add a tag describing the correct genre. Moreover, tag recommendations can be useful to promote the use of a particular form of a concept (e.g., suggesting the tag `street-music` instead of `streetmusic`), to prevent misspellings when introducing tags and, in general, to facilitate the tagging process (Sood et al., 2007; Jäschke et al., 2007; Wang et al., 2012). As a result, many authors hypothesize that tag recommendation

systems have a positive impact on the folksonomy of a tagging system, improving the quality of resource annotations (Jäschke et al., 2012; Wang et al., 2012) and the coherence and convergence of the vocabulary of the folksonomy (Golder & Huberman, 2006; Marlow et al., 2006; Jäschke et al., 2007; Sood et al., 2007; Zangerle et al., 2011). This is also one of the aspects that we investigate in this thesis.

Nevertheless, we would like to point out that tag recommendation systems could also be expected to have a negative impact on the folksonomy of a tagging system. Even though we are not aware of such claims in the tagging literature, we could hypothesize that an excess of homogenisation in resource annotations would result in a loss of their informational value, making it hard to distinguish among resources just by looking at their annotations. However, considering that non-uniformity of resource annotations is typically listed as an important issue of tagging systems (Spiteri, 2007), the degree of homogenisation that would be required in order to effectively reduce informational value seems intuitively far from its current state. As another possible negative aspect, it could be argued that providing tag suggestions could make it easier for users to poorly annotate their content by, for example, randomly choosing tags from the list of recommendations. Nonetheless, this kind of resource annotations could also be performed without the use of a tag recommendation system. Hence, this potential problem seems to be more related with the motivations that users have when annotating content (see 2.2.2). For all these reasons, in this thesis we do not evaluate the hypothetical negative impact of tag recommendation systems.

1.4 Online multimedia sharing

Multimedia sharing is one of the areas in which the social web has experienced the biggest and quickest growth (Smith, 2009). It can be roughly divided into video, image and audio sharing. Just to name a few examples, in every minute, 100 hours of video are uploaded to YouTube (The YouTube Team, 2013), 2,400 photos are uploaded to Flickr (Jeffries, 2013), and 12 hours of music are uploaded to SoundCloud (Wahlforss & Eric, 2013). The intent with which users upload and share multimedia content can vary widely, but we can identify some general patters according to the usage that the uploaders may expect of the contributed content. On the one side, we can identify those contents that are meant to be accessed and consumed through the online sharing platform itself. In that case, users might upload a photo, a video or a music track and expect other users to consume it *in-place*. Hence, the *end use* of the resource is its online consumption. For example, someone may upload photos of an event to a photo sharing site so that other participants of that event can have access to the photos, or a musical artist can upload a music album to a music sharing site so that other users can listen to it. On the other

side, there is an additional type of uploaded content which is meant to be reused outside the sharing platform where it is hosted. Here, the consumption in the sharing platform does not represent an end use per se. Some examples of this situation include sharing sound effects that can be later used in video games, drum loops in music compositions, video backgrounds or transitions to be used in audiovisual installations, or images to be used in collages or as a desktop wallpaper. These latter cases of multimedia sharing particularly support the *read/write culture* concept introduced by Lessing (2008), in which users are both consumers and producers of content that is easily shared and reused through the internet (Wikipedia, 2014a). Conversely, the case in which content is only consumed in the sharing platform is closer to the *read only culture*, in which content is shared but not reused.

In both cases, the challenges for describing, indexing and retrieving uploaded content prevail. However, considering the previous ideas, it seems plausible that multimedia sharing for the read/write culture can pose more complex scenarios. In the read/write culture, users may need more sophisticated and specialised ways of accessing online resources that fit their particular requirements. For example, a user might need background images with a specific set of colours that fits some other images, or she might need a drum loop with a specific tempo and playing style. In that case, description and indexing processes become even more essential to provide proper content access in multimedia sharing sites.

1.5 Online sound sharing

Among the different kinds of multimedia that are commonly shared online, the case of sound sharing is particularly interesting. By *sounds* (or *audio clips*), we understand any kind of audio material like sound effects, environmental recordings or even building blocks for musical compositions, but not music tracks in the traditional sense of "finished" compositions or songs. Users searching for content in sound sharing sites might be looking for audio clips with very specific and detailed characteristics that can be represented by a wide range of audio properties. For example, a user might be searching for the sound of an opening door with a particular duration, size and material of the door, or a user might be searching for the sound of a melody being played by a particular instrument with a specific tonality, tempo and mood. Being able to successfully retrieve such specific content resources poses a very challenging problem to both the users and the sharing platform. Another relevant aspect of sound sharing is that the assessment of the results returned by a search engine of a sound sharing site requires the time to listen to them, and can not be done as instantly as it could be done with the search results of, for example, a photo sharing site. From this point of view, the cost of iterating over several queries in order to find the desired resource is higher for sounds (and also for music

Site	URL	# Sounds	Introduced
Sound Dogs	http://www.sounddogs.com	670,000	1997
Freesound	http://www.freesound.org	230,000	2005
Sound Snap	http://www.soundsnap.com	160,000	2007
SFX Source	http://www.sfxsource.com	140,000	2007
Free Sound Effects	http://www.freesoundeffects.com	100,000	2012

Table 1.1: Most important sound sharing sites according to their estimated number of shared sounds. The data shown in this table is approximate and gathered from various sources such as the "About" or "Frequently Asked Questions" sections of the sites, copyright notes, and the "Wayback Machine" of the Internet Archive[17].

and video) than for images. This increases the importance of the description, indexing and retrieval challenges. Finally, it is also important to note that most of the existing multimedia sharing related research is devoted to either photo, video or music sharing, but sound sharing is generally underattended.

Despite online sound sharing not being as widespread as video, photo or music sharing, there exist a relatively large number of sharing platforms solely focused on sounds that host large amounts of content. Table 1.1 shows a list of the most important online sound sharing sites according to an estimated number of hosted sounds. In general, these sites host a mixture of content generated and annotated by hired professional sound designers, and a small amount of content generated and annotated by the users of the site. Hence, these are consumer-oriented sites which do not feature a strong user community and do not provide a standard mechanism for uploading content. Even though a small amount of the hosted audio content can be freely accessed, most of it requires the payment of a fee in order to be downloaded. However, of special relevance to the present thesis is the case of Freesound. Unlike the other sound sharing sites listed in Table 1.1, all the content in Freesound is uploaded and annotated by its community of users, and is released under open Creative Commons licences that do not require the payment of any fee for its use. This makes Freesound the site whose nature is closer to the phenomenon of multimedia sharing that can be observed in sites like Flickr or YouTube, and to the aforementioned philosophy of the read/write culture and easy content sharing. There exist other similar sites like Looperman[15] or ccMixter[16] which feature comparable characteristics in terms of user community and openness of the content, but these are strongly music oriented, not entirely fitting in our definition of sound sharing, and thus not included in Table 1.1.

Freesound is an online sound sharing site that was started in the research

[15] http://www.looperman.com.
[16] http://www.ccmixter.org.
[17] http://www.archive.org/web/

group where this thesis has been carried out. It was launched in 2005 with the goal of creating an open licensed database of sounds that could be used for research purposes. Over the years, it has become a sound sharing site of reference, featuring an average of 37,000 unique visits per day, 160 newly uploaded sounds per day, and ranking the 10,387th most visited worldwide web site in the Alexa ranking (significantly above the other sound sharing sites listed in Table 1.1)[18]. Freesound sounds are mainly described using textual descriptions and tags provided by the users that upload them. At the time of this writing, Freesound features a folksonomy with 1,670,000 tag applications relating 77,000 tags, 230,000 sounds, and 12,000 users (only considering users that have uploaded and annotated at least one sound).

Freesound closely fits in the sharing paradigm of user generated content and, as a sound sharing platform, faces all the challenges that we have described above and in the previous sections. Hence, Freesound's nature and the fact that it is still currently developed and maintained in the research group, makes it an ideal use case for the purposes of this thesis. Through Freesound, we have been able to implement and evaluate our recommendation methods in the real world, and analyse the impact that tag recommendation has had in the large-scale scenario of Freesound. To our knowledge, this is the first work that has been able to perform such analyses in such optimal conditions.

1.6 Objectives and outline of the thesis

In the previous sections we have explained the motivations, described the context and introduced the focus of our thesis. In accordance with all that has been said, the main goal of this thesis is to contribute to advancing the state of the art in folksonomy-based tag recommendation systems by proposing and thoroughly evaluating several methods and their impact in a real-world sound sharing scenario. Even though this thesis is focused on the particular case of sound sharing, the work we present strives for generalization to other multimedia domains, and thus can be of interest to researchers working in other fields. Fig. 1.4 shows a conceptual organisation of some of the chapters of this thesis according to the domain-specificness and the amount of knowledge embedded in the tag recommendation methods described therein. What follows is a brief description of the structure of the thesis along with the specific goals and achievements that are reported in each chapter.

In Chapter 2, we provide a comprehensive literature review centred around tagging systems and the task of tag recommendation. We start by describing what tagging systems are and the different categorisations of tags and tag-

[18]These statistics have been computed considering Freesound data from 3 August 2013 to 2 August 2014. Alexa's ranking information was retrieved on 2 August 2014 (`http://www.alexa.com/siteinfo/freesound.org`). More information on Freesound statistics can be found in Appendix A.

Figure 1.4: Conceptual organisation of Chapters 3 to 6 of this thesis according to the characteristics of the tag recommendation methods described therein.

ging systems. We also look at the potential motivations that users have when tagging resources, and at the typical problems of tagging systems. Then, we specifically focus on tag recommendation and provide a review of several methods that have been proposed in the literature. In addition, we describe which are the typical strategies followed in the literature for evaluating tag recommendation methods. Finally, we discuss about the potential impact of these methods in online sharing platforms.

In Chapter 3, we describe a generic scheme for folksonomy-based tag recommendation. This consists of three steps for which we provide several strategies. By combining these alternative strategies, we define several tag recommendation methods. These methods are systematically evaluated and compared to other state of the art folksonomy-based tag recommendation methods through a tag prediction task and using data from Freesound and Flickr. The goal of this chapter is, on the one hand, the definition of a common scheme and conceptualisation of the tag recommendation task that will be used in in subsequent chapters. On the other hand, in this chapter we compare our proposed methods with state of the art alternatives. Our evaluation shows that the proposed methods can successfully generate tag recommendations both in the image and audio domains, performing significantly above the other evaluated methods.

Chapter 4 extends the best performing tag recommendation method presented in the previous chapter by including a new step in which the resources being annotated are automatically classified into five broad audio categories. Using this classification, the system is able to tailor tag recommendations to every particular audio category. The recommendation system described in this chapter takes advantage of simple knowledge, specific to the audio domain, embedded in the audio classifier. It is evaluated against the best performing method presented in the previous chapter, and against two random baselines through an online user experiment with 190 participants carried out in the context of

Freesound. Furthermore, a complementary evaluation is also performed following the same evaluation strategy as in the previous chapter. The objective of this chapter is to evaluate whether the introduction of simple domain-specific knowledge in the form of resource categories can improve the usefulness of the tag recommendations generated by our previous method. Results show that the extended recommendation method represents an improvement over the previous method.

In Chapter 5, we analyse the impact of the tag recommendation method described in the previous chapter after we introduced it in the real-world tagging system of Freesound. The goal of this chapter is the assessment of several hypotheses that have been made in the tagging literature regarding the impact that tag recommendation methods could have on the folksonomies of tagging systems. To the best of our knowledge, this is the first time that such an analysis has been made. For this, we propose a series of evaluation metrics to illustrate different aspects of these hypotheses, and compute such metrics over data gathered during three months of activity after the deployment of the tag recommendation system, and over data from the previous two and a half years. The analysis reveals that the tag recommendation system effectively contributes to the vocabulary convergence of the Freesound folksonomy, partially contributes to an improvement of annotation quality, but does not seem to significantly reduce the cost of the tagging process.

In Chapter 6, we explore a new perspective for folksonomy-based tag recommendation in which we introduce more domain-specific knowledge modelled with an ontology. We describe an ontology which embeds information about audio categories, tag categories and their relations. We describe a prototype for a tag recommendation system which makes extensive use of the ontology to provide the recommendations and evaluate it with two online experiments. Our particular goal in this chapter is to explore whether we can build tag recommendation systems that, by taking more advantage of domain-specific and structured knowledge, can help users in generating better quality resource annotations. Results show that using the tag recommendation prototype we describe, users can effectively generate better quality resource annotations. Nevertheless, we also observe that several improvements should be made before deploying such a system in a real-world scenario.

At the end of each chapter, we include a focused discussion about the relevant results and conclusions. We conclude this thesis in Chapter 7 with a summary of our work, our main conclusions, and with a discussion about future perspectives of sound sharing, tag recommendation, and tagging systems in general.

CHAPTER 2

Literature review

2.1 Introduction

The literature review presented in this chapter is divided into two parts. Firstly, we summarise existing work on the definition and characterisation of tagging systems. We describe what a tagging system is, how different authors have proposed to categorize tagging systems, the motivations that users have when annotating content and the different types of tags resulting from these annotations. Additionally, we discuss problems that are typically found in tagging systems and highlight some of the solutions that are commonly proposed. Secondly, we focus on existing specific literature about tag recommendation systems. We outline the different approaches that have been proposed for tag recommendation, describe how these approaches are normally evaluated, and finally summarise research about the impact that tag recommendation is expected to have on the folksonomies of tagging systems.

2.2 Tagging systems

Tagging systems systems have been well studied since the popularisation of tags in online sharing platforms and the social web in general. From a broad perspective, tagging is the process of assigning tags to content resources. Considering that this can be done manually by users or automatically by machines, tagging approaches can be divided into *manual* and *automatic* tagging (Wang et al., 2012). The focus of this thesis is on manual tagging systems. Manual tagging systems can provide different levels of assistance during the tagging process (e.g., the system can suggest popular tags to the user) or provide no assistance at all, but in both cases users have the final decision on whether or not to assign a given tag to a content resource. Fig. 2.1 shows the manual tagging interface of three online multimedia sharing sites. Contrastingly, automatic tagging systems are designed to automatically add tags to resources without the need (or very little need) of human intervention. Even though in this thesis

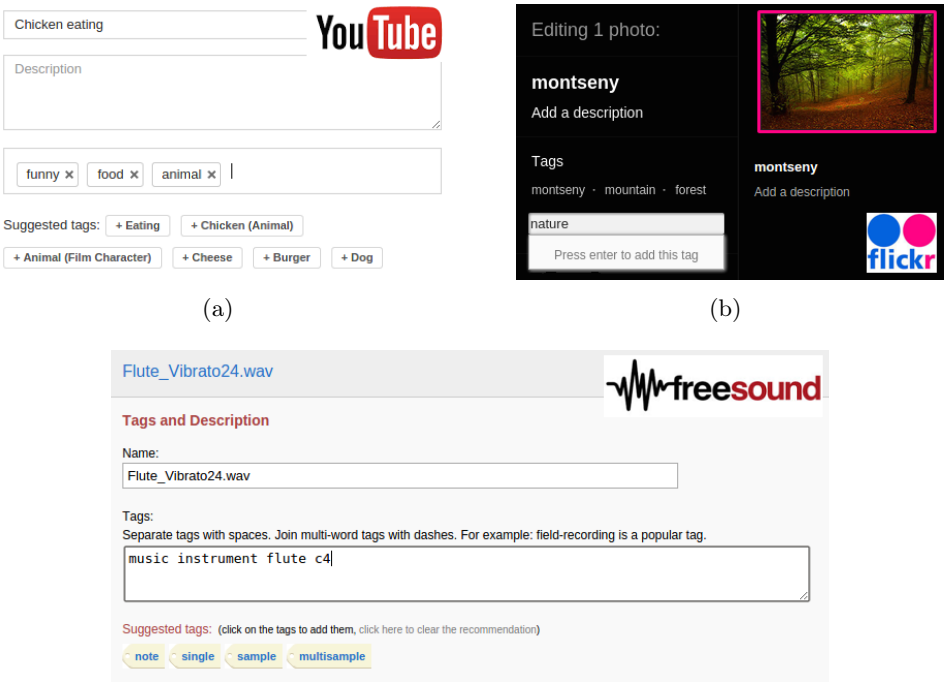


Figure 2.1: Examples of manual tagging system interfaces for (a) YouTube, (b) Flickr, and (c) Freesound. In this examples, both YouTube and Freesound annotation interfaces provide a number of suggested tags, whereas the Flickr interface does not.

we will not deal with automatic tagging systems, these are closely related to the tag recommendation systems that will be discussed later in this chapter, and thus we give an overview of them in Sec. 2.3.

In tagging systems, the individual tags that are assigned to a content resource typically represent annotations about different information dimensions or facets which, when considering the whole tagline, conform a description of the resource. For example, a tagline for a sound representing a recording of the ambience in "La Boqueria" market in Barcelona, might include the following tags: {`coins`, `barcelona`, `boqueria`, `atmosphere`, `market`, `noise`, `people`}[19]. In this case, the tags `barcelona` and `boqueria` indicate a specific location of the recording, `market` conveys information about a generic place, `people` and `coins` list sound-producing elements appearing in the recording, `noise` describes an acoustic property of the recording, and `atmosphere` provides a possible generic classification of the type of audio. Other tags could be added, for example, to indicate the recording device used to capture the audio, or the specific time when the recording was made. These tags are not introduced with a formal

[19]This example is actually taken from a sound shared in Freesound. It can be accessed in the following URL: `http://www.freesound.org/people/helenacm/sounds/101792/`.

structure, and users can not indicate explicit relations among them. The tags in the tagline, and by extension the aggregate of all distinct tags that form the folksonomy of a tagging system, are organised in a flat namespace, meaning that there is no explicit nor predefined hierarchy in which tags can be projected. Hence, the tagging activity is considered a categorisation process rather than a classification process, where the different tags convey information about possibly overlapping facets without formal structure (Jacob, 2004; Halpin et al., 2006). Conversely, formal classification systems organise items in unambiguous and exclusive concept hierarchies (or taxonomies) with a predefined and controlled vocabulary (Golder & Huberman, 2006). It has been argued that the use of tags instead of taxonomies with controlled vocabularies to annotate resources has two main advantages. On the one hand, tags can better adapt to the annotation requirements of resources which might not be contemplated in controlled vocabularies. This brings a lot of flexibility as users can intuitively introduce previously non-existing tags to annotate unexpected content properties (Mathes, 2004; Shirky, 2005; Quintarelli, 2005; Halpin et al., 2006; Sen et al., 2006; Fichter, 2006; Wu et al., 2006; Macgregor & Mcculloch, 2006). On the other hand, when using tagging systems, users do not need to learn and understand the specific terms and organisation of a taxonomy, reducing the cognitive load of the annotation process and smoothing its learning curve (Quintarelli, 2005; Shirky, 2005; Fichter, 2006). This is one of the reasons why tagging systems have become widespread in the social web and online sharing platforms in particular (Cattuto, 2006). However, tagging systems do exhibit significant disadvantages when compared with formal classification systems (see Sec. 2.2.4).

2.2.1 Types of tagging systems

Marlow et al. (2006) propose a categorisation for tagging systems according to seven dimensions, and discuss the implications of design choices that can be made for each dimension. In Table 2.1 we show some of the most popular online sharing platforms categorised according to their tagging system's characteristics and the dimensions proposed by Marlow et al. (2006). The first dimension, "tagging rights", indicates which users of the tagging system can assign tags to resources. In some systems, only the owner of a resource can assign tags to it, while in other systems resources can be tagged by any user. As we have seen in the introductory chapter, such a design choice leads to the emergence of either a narrow or broad folksonomy, respectively. The second categorisation dimension, "tagging support", divides tagging systems according to the level of assistance given to the user during the tagging process. On the one hand, there are blind systems in which users are given no particular support during the tagging process. On the other hand, there are systems that feature mechanisms with different levels of complexity to suggest tags to users or guide the annotation process. Authors suggest that non-blind tagging

systems reinforce the convergence of the vocabulary in the folksonomy (see Sec. 2.3.5). The third categorisation dimension, "aggregation model", specifies whether the tags assigned by different users to a single resource are considered as a set of unique tags or as a union of all tags (bag of tags). Hence, this dimension is only applicable to those tagging systems where resources can be tagged by multiple users (broad folksonomies). The different aggregation models constrain the available tagging statistics for each resource. The fourth dimension proposed by Marlow et al. (2006) is the "object type", and defines the nature of the resources being tagged (e.g., web pages, bibliographic material, audio, video, images, etc.). The authors suggest that the type of resource has numerous implications on the resulting tags and folksonomy, hypothesising that resources without a direct textual representation (e.g., multimedia resources) require different kinds of annotations. The fifth dimension, "source of material", separates tagging systems in those which the content to be tagged is supplied by users of the platform (typically user generated content or content from external sources like web pages), or by the platform itself. Authors suggest that the incentives that users have for tagging the content will be different depending on the source of the material. Finally, the sixth and seventh dimensions are "resource connectivity" and "social connectivity", and indicate whether the resources or users of the system are explicitly organised in groups based on similarity, shared interests or any other aspects. Marlow et al. (2006) suggest that explicit social or resource connectivity can contribute to the emergence of localised folksonomies for the defined groups, yielding particular tagging behaviours which are distinguishable from the global behaviour. Sen et al. (2006) also proposed a categorisation of tagging systems that is very similar to the above described. Although it only consists of four dimensions ("tag scope", "item ownership", "tag selection" and "tag sharing"), almost direct equivalences can be drawn.

2.2.2 User motivations for tagging

Several authors have studied the motivations or incentives that users have when tagging online resources (Mathes, 2004; Marlow et al., 2006; Golder & Huberman, 2006; Sen et al., 2006; Xu et al., 2006; Ames & Naaman, 2007; Gupta et al., 2010). Marlow et al. (2006) propose a high-level categorisation of potential user motivations in two categories: "organisational" and "social". The "organisational" category includes tagging activities motivated by the aim of bringing a particular structure and organisation to contributed resources. In this case, users might develop particular tagging patterns and also adopt common patterns observed in other users of the tagging system. Tags under the "organisational" category should, in general, ease the future retrieval of the resources being tagged. Tagging activities under the "social" category include the listing of user opinions and other communicative aspects that allow users to express themselves. Besides these two broad categories, Marlow et al.

	Flickr	YouTube	Soundcloud	Freesound	Delicious	Bibsonomy	CiteULike
Tagging rights	Owner	Owner	Owner	Owner	Everyone	Everyone	Everyone
Tagging support	Blind	Tag rec.	Auto-completion	Tag rec.	Tag rec.	Tag rec.	Blind
Object type	Photos	Video	Music and sounds	Sounds	Bookmarks	Biblio. ref., bookmarks	Biblio. ref.
Source of material	User-generated	User-generated	User-generated	User-generated	User-provided	User-provided, user-generated	User-provided, user-generated
Resource connectivity	Albums	Categories	Sets	Packs	-	-	-
Social connectivity	Groups, Followers	Channels, Followers	Groups, Followers	Followers	Followers	Groups	Groups, Followers

Table 2.1: Popular online sharing sites categorised in the dimensions defined in Marlow et al. (2006). As the functionalities of some of these systems changed over time, we list here its characteristics in their state at the time of this writing. Note that the "aggregation model" dimension is not included in the table because it is typically not available.

(2006) also describe several more specific potential motivations for tagging resources. Gupta et al. (2010) summarise these more specific categories along with others found in the literature (Mathes, 2004; Golder & Huberman, 2006; Xu et al., 2006; Sen et al., 2006; Ames & Naaman, 2007), and finally enumerate ten specific and non-exclusive categories which exemplify possible user motivations for tagging resources. Table 2.2 lists these categories and provides a brief explanation for each one.

Besides these categorisations, of particular interest is the work by Ames & Naaman (2007), in which an empirical tagging study is carried out with participants being interviewed about their motivations when tagging pictures to be uploaded to Flickr. In this study, the authors classify the responses provided by the interviewed participants in similar categories as those listed in Table 2.2. The observed motivations are, basically, those related to expressing user's opinions or relation with the photos to other known users of the system such as family or friends (roughly corresponding to "self presentation", "opinion expression" and "social signalling" categories of Table 2.2). Moreover, motivations related to the organisation of the resources (for easing future retrieval to both the owners of the photos and to other users of the system) are also repeatedly reported. Of particular relevance is the fact that, according to the interviews, participants are not aware of all potential motivations and benefits of tagging systems when annotating resources, and the decisions of which tags to choose are taken in an intuitive way rather than in an informed fashion. In relation to this, in a study by Marlow et al. (2006), also with Flickr tagging data, authors suggest that the more resources a user has tagged, the more aware the user is

Name	Description
Future Retrieval	Tags that users add to facilitate the future retrieval of the annotated resources. These tags can describe particular properties of a resource which are relevant to a broad audience and can be effectively used for indexing purposes. However, future retrieval tags can also include information which is only relevant to the user performing the annotation such as a tag like `toread`, which is typically used as a future reminder for the annotator.
Contribution and Sharing	These are tags that serve the purpose of categorising resources into rather common concepts and facilitate in this way future retrieval for other users of the platform.
Attract Attention	Some users choose to add popular tags when annotating their own resources to deliberately increase their reachability. This particular case can have a negative impact in the quality of the annotations if popular tags are chosen that have no relation with the actual content.
Play and Competition	Some tagging systems feature interfaces in which the annotation process is presented as an entertaining activity where users can, for instance, cooperate on annotating a resource. These systems are typically known as "games with a purpose" (Von Ahn, 2006). The most well known example of a game with a purpose in the tagging field is the ESP game (Von Ahn & Dabbish, 2004), in which pairs of users need to concurrently annotate a particular resource and are rewarded a number of points in accordance with their agreement in the chosen tags.
Self Presentation	These are tags that bear some aspect of the identity of the user that annotates a resource. Marlow et al. (2006) show, as an example of this kind of tags, the tag `seen live`, which sets a personal relation between the annotator and the resource.
Opinion Expression	Tags that users add with the purpose of expressing their subjective judgement or opinion about a resource.
Task Organisation	Similarly to some tags that could be in the future retrieval category, task organisation tags are used for organising resources through associated tasks that a particular user relates with the resource (e.g., `todo`, `toread`).
Social Signalling	Tags can be chosen to convey contextual information about a resource. For example, the name of the event in which a photo has been taken. In this case, users might be motivated by communicating their presence at that event.
Money	In some cases, users are being paid to annotate resources, typically through the use of platforms like the Amazon Mechanical Turk (`http://www.mturk.com/mturk/welcome`).
Technological Ease	Users can be also motivated by the ease of use of tagging systems (and sharing platforms in general) to annotate resources, so that the easier the tagging process is, the more likely users will be to annotate resources.

Table 2.2: Potential tagging motivations listed by Gupta et al. (2010).

about the relevance of chosen tags and annotation quality. In other words, as users learn to tag, their motivations change.

2.2.3 Types of tags

In the previous section we have seen how some authors categorise tags in terms of potential user motivations or incentives. Another dimension in which tags can be categorised is on the basis of the kind of information that these convey about resources. In that direction, several authors have proposed different, but highly related categorisations which are summarised in Table 2.3 and which are generic enough to be applied in tagging systems of different domains (Golder & Huberman, 2006; Sen et al., 2006; Xu et al., 2006; Bischoff et al., 2008; Gupta et al., 2010; Cantador et al., 2011). Among these categorisations featuring broader tag categories (upper rows in Table 2.3), we find particularly meaningful the four-class categorisation proposed by Cantador et al. (2011). Tags under "content-based" category describe the objects and qualities of a resource (e.g., content-based tags might enumerate the musical instruments that are present in an audio resource or its music genre). Tags under "context-based" category provide information about the context in which the resource was created (e.g., the location where a photo was taken or the time of the year in which a video was recorded). "Subjective" tags are those which express personal opinions that the tagger has about a resource at hand, such as quality judgements or mood annotations. Finally, "organisational" tags annotate resources with information that is, a priori, only useful for the annotator of the resource such as reminders related to the resource or self-referencing comments. As can be seen in Table 2.3, the other proposed broad categorisations are very similar (Sen et al., 2006; Xu et al., 2006). Contrastingly, categorisations proposed by Golder & Huberman (2006), Bischoff et al. (2008), and Gupta et al. (2010), include more fine-grained categories that can be easily understood as subdivisions of those broader categories mentioned above.

Empirical research on the categorisation of tags has also been performed. Simons (2008) analyses the Flickr tagcloud and performs a manual classification of the tags into a list of categories crafted to fit the data. By mapping these tailored categories to the broader categories proposed by Cantador et al. (2011), it can be seen that around 66% of tags belong to either "content-based" or "context-based" categories. The remaining 33% can be classified as "subjective" or "organisational" tags. Similarly, Bischoff et al. (2008) perform an analysis on the distribution of tags among their proposed categories using data from Delicious, Last.fm and Flickr. After a manual classification of a sample of all used tags, 55% of the tags belong to either "topic", "type" or "location" categories, which are related to the broader "content-based" and "context-based" categories defined by Cantador et al. (2011). Furthermore, Cantador et al. (2011) propose a rule-based method for automatically classifying tags into their proposed four broad categories. The method is based on the use of natural

Sen et al. (2006)	Factual[a]					Subjective	Personal[a]		
Cantador et al. (2011)	Content-based			Context-based		Subjective	Organisational		
Xu et al. (2006)	Content-based	Attribute		Context-based		Subjective	Organisational		
Golder & Huberman (2006)	What or who it is about	What is	Who owns it	Refining other categories		Identifying qualities or characteristics	Task organisation	Self reference	
Bischoff et al. (2008)	Topic	Type	Author or owner	Time	Location	Opinions and qualities	Usage context	Self reference	
Gupta et al. (2010)	Content-based	Attribute	Ownership	Context		Qualities and characteristics	Organisational and Purpose	Self reference	

[a] These categories are also listed in Gupta et al. (2010).

Table 2.3: Types of tags according to the kind of information conveyed about resources. This table is adapted and extended from the works of Cantador et al. (2011) and Bischoff et al. (2008).

language processing techniques and YAGO[20], an external knowledge base. By performing a part-of-speech analysis of the tags and matching them to concepts of the YAGO knowledge base, the authors are able to determine to which category a tag belongs. Using data collected from Flickr, the authors performed the classification and found that, among those tags whose category could be predicted, 64% are considered to be either "content-based" or "context-based" tags, while the others belong to "subjective" or "organisational" categories. As can be observed, these results are consistent with those reported by Simons (2008) and Bischoff et al. (2008).

By comparing Tables 2.2 and 2.3, it can be easily seen that tag types and tagging motivations are tightly coupled. We explained before in Sec. 2.2 that Ames & Naaman (2007) found that users are more motivated for introducing tags expressing subjective opinions and self-references than to introduce tags describing the nature of the resources for their organisation. Considering this observation, we would expect to find more tags corresponding to the "subjective" and "organisational" categories rather than to the "content-based" and "context-based" categories, which is not what has been observed by Simons (2008), Bischoff et al. (2008), and Cantador et al. (2011). Interestingly, among the resulting types of tags and motivations, not all of them are equally suitable for generating useful metadata for indexing the content of online sharing platforms. In the case of systems featuring narrow folksonomies such as Flickr or Freesound, those tags that convey information which is meaningful not only to the owners of a resource but also to the other users of the platform, are crucial in order to successfully index content. Thus, according to the categorisation proposed by Cantador et al. (2011), the presence of "content-based" and "context-based" tags is more desirable than "subjective" and "organisational" tag types. Conversely, in broad folksonomies such as Declious or Last.fm, "subjective" and "organisational" tags can be as important as "content-based" and "context-based". This is because in these systems users mainly tag for their self-organisation, and the used tagging conventions do not necessarily need to be meaningful to other users of the platform (De Meo et al., 2013).

2.2.4 Tagging systems' problems and solutions

We have seen that the flexibility provided by tagging systems typically carries a number of well known problems which limit the possibilities of indexing, searching and browsing in sharing platforms (Golder & Huberman, 2006; Halpin et al., 2006; Guy & Tonkin, 2006). A very common problem is the presence of tags with typographical errors and tags formed with several concatenated words. Guy & Tonkin (2006) found that 40% of Flickr tags and 28% of Delicious tags contain misspellings or compound words that could not be

[20] `http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/`.

mapped into a dictionary. These tags become less relevant, as a tagging system most probably treats misspelled versions of words as different tags. Moreover, word concatenation easily leads to different variations of a single concept. In addition to that, it is very common that a single tag might have several different meanings (polysemy), and thus some users might employ it thinking of one meaning and some other users might employ it thinking of another meaning. Without a successful method for disambiguation, this results in making relations between resources which are semantically not meaningful. For example, search results might display resources related to different meanings of the query terms. Conversely, it is also quite common that several tags refer to a single concept (synonymy) and users employ them indistinctly. In that case, some relevant resources might be left out of the results of a query because systems are not generally aware of synonymy relations. Polysemy and synonymy problems have been empirically evaluated by Spiteri (2007), analysing data of folksonomies gathered from Delicious, Furl and Technorati[21], where it was found that between 12% and 22% of the tags potentially feature these kind of problems.

Another common problem of tagging systems is the use of tags that are only relevant for a specific user or the use of tagging conventions which are only known by particular users. These kind of tags convey information which is not generally useful to the community, and therefore their organisational value is limited. For example, a user might annotate resources using very particular tags whose meaning is not known to other users, and these annotations might appear to be totally unrelated to the resources. Kennedy et al. (2006) found that, given a Flickr image, only 50% of the tags can actually be easily related to the content of the image or even to the image at all. That can particularly become a problem in tagging systems with narrow folksonomies, in which tags that have a shared meaning for the community should be reinforced to improve indexing, searching and browsing possibilities (Guy & Tonkin, 2006). Furthermore, a typical problem of tagging systems is the *lack of tags*. Some studies show that user provided tags tend to be incomplete. As an example, Sigurbjörnsson & Zwol (2008) show that 50% of images in Flickr have less than four tags, and Zhao et al. (2010) show that YouTtube videos have an average of five tags, which means that annotations are not very comprehensive.

On an even higher level, different tagging styles can also create a problem for the tagging system if there are no signs of consensus. If a common vocabulary is not shared among users, the informational value of tags is lessened and resources become less reachable (see Sec. 2.3.5). Furthermore, the co-existence of different languages in a single tagging system can also become an obvious problem (Halpin et al., 2006). Overall, most of these issues are inherent to the

[21] Furl is a no longer existing online sharing platform. In it, users shared web bookmarks (similarly to Delicious). Technorati is currently a publisher advertising platform, but used to be a blog tracking site (`http://www.technorati.com/`).

vocabulary problem (Furnas et al., 1987), caused by the lack of well defined tagging guidelines and the different rationales that users can apply during the tagging process. Nevertheless, the design and functionalities of tagging systems potentially have a big influence on the tagging behaviour when annotating resources, and this could be shaped so that typical tagging problems can be lessened (Wang et al., 2012).

In order to mitigate some of the above mentioned tagging systems' problems, many authors have focused on approaches in which manual tagging is combined with computer algorithms to try to "optimise" the taglines introduced by users. Some of these approaches are meant to be used at the time when users annotate newly uploaded resources, but others are focused on improving already existing annotations. Wang et al. (2012) introduce the concept of *assistive tagging* to refer to these approaches, and propose a classification into three groups. The first group, "tagging with data selection and organisation", includes these approaches in which tagging systems automatically detect already existing resources which are poorly described and ask users to annotate them. In this case, many users can contribute to improving the annotations of a pool of selected resources, and the system can prioritise which content should be annotated first (Huang et al., 2008; Wang & Hua, 2011). A similar idea consists of the organisation of data to be annotated in clusters, which then become the smallest unit of tagging. Tags assigned to particular clusters are propagated to all resources of that cluster. This approach has been mainly applied in photo tagging, with clusters formed on the basis of face recognition algorithms (Suh & Bederson, 2004; Cui et al., 2007; Tian et al., 2007), or on the basis of sub regions of an image (Tang et al., 2010). In the latter case, it is interesting that by clustering photos according to smaller regions and then annotating the regions, the tags can be applied to many photos at once. These methods are oriented to batch tagging of resources which, in the context of an online sharing site, does not necessarily need to be performed at upload time by the authors of the content. Instead, it can be performed in a collaborative fashion by other users of the platform.

The second group of assistive tagging strategies, "tag processing", includes those approaches in which existing annotated resources are post-processed in order to automatically correct or refine the descriptions. For example, in photo tagging systems, images can be segmented and machine learning algorithms can be trained to learn the mapping of introduced tags with particular regions, and then propagate the tags to other images with similar regions (Liu et al., 2010b; Feng et al., 2010; Liu et al., 2011). Similar ideas, applied at a temporal level rather than at the region level, have been applied to video tagging (Ulges et al., 2008). Furthermore, systems can be trained to compute a relevance score for the tags assigned to a resource based on the analysis of its content, and filter in this way tags with low scores (Liu et al., 2009; Li et al., 2009; Chen et al., 2010a; Fan et al., 2010). Another approach for post-processing user

provided annotations is the use of knowledge bases like WordNet (Miller, 1995) to extend annotations with synonyms and hypernyms or filter out tags which are unrelated according to the knowledge base (Liu et al., 2010a). Finally, the third group of assistive tagging strategies identified by Wang et al. (2012), "tag recommendation", consists of approaches in which tagging systems suggest tags to users during the annotation process, thus potentially shaping users' tagging behaviour. These systems are the main topic of the present thesis, and are discussed in the following section.

2.3 Tag recommendation

The existing literature on tag recommendation systems is generally focused on recommendation systems for image or social bookmarking sharing sites. However, highly related to tag recommendation systems are automatic tagging systems. In essence, automatic tagging systems and tag recommendation systems share their main goal: generating a set of relevant tags for a given resource. Hence, a lot of the techniques described in the literature are applicable to the two kinds of systems. In this section, we summarise a number of approaches that, either being more focused on tag recommendation or automatic tagging, are of relevance to contextualise the work described in this thesis. Nevertheless, we put our focus on the systems designed for the task of tag recommendation.

Tag recommendation systems can be classified according to the main source of information that is used in the recommendation process. In general, approaches can be separated in *i)* systems based on the content analysis of the resources, *ii)* systems based on the folksonomy of a tagging system, and *iii)* systems based on contextual data (Wang et al., 2012). In the following sections, we review existing literature on each one of these approaches. Table 2.4 shows a summary of all approaches reviewed in these sections.

2.3.1 Based on content analysis

Content-based tag recommendation systems take advantage of the analysis of the content of resources in order to provide a set of recommended tags. For example, given an image, a content-based tag recommendation system leverages the pixel data to extract features like colour, texture and shape that can be used to derive a set of potentially relevant tags. These approaches are marked with the abbreviation "CO" in the "type of approach" column of Table 2.4. One way in which relevant tags can be recommended after the extraction of low-level features is the use of machine learning techniques to build a model for every potential tag to be recommended. This typically implies to learn the joint probabilities between content features and the presence of particular tags (marked with the abbreviation "MOD" in the "based on" column of Table 2.4).

	Focus	Type of resource	Type of approach	Based on	Sharing platform	Evaluation method	Dataset size	Evaluation measures
Barnard et al. (2003)	Auto Tag.	Image	CO	IF, MOD	-	Prediction	9,000	Increase of P
Barrington et al. (2007)	Auto Tag.	Sound	CO	AF, MOD	-	Search	1,305	P, R, F
Jäschke et al. (2007)	Tag. rec.	Bookmarks, Music	FO	CF, FK, PE	Bibsonomy, Delicious, Last.fm	Prediction	361 74,854 1,853	P, R, F
Sood et al. (2007)	Tag. rec.	Blog posts	CT	TF, SIM	Technorati	User assess. Prediction	225	P, R, F
Anderson et al. (2008)	Tag. rec.	Image	FO, CO	CC, IF, MOD	Flickr	Prediction	924	MRR, S, P
Chen et al. (2008)	Tag. rec.	Image	CO	IF, MOD	Flickr	User assess.	930	[a]
Garg & Weber (2008)	Tag. rec.	Image	FO	CC, PE	Flickr	Prediction	50M	MRR, S, P
Li & Wang (2008)	Auto Tag.	Image	CO	IF, MOD	Flickr	User assess. Prediction	5,411 47,000	P, R, F, S
Lipczak (2008)	Tag. rec.	Bookmarks	FO	CC, PE	Bibsonomy	Prediction	274,139	P, R, F
Naaman & Nair (2008)	Tag. rec.	Image	CT	GPS	Flickr	Prototype	100,000	[b]
Sigurbjörnsson & Zwol (2008)	Tag. rec.	Image	FO	CC	Flickr	User assess.	331	MRR, S, P
Song et al. (2008)	Tag. rec.	Bookmarks	FO	CC, TCL	Bibsonomy	Prediction	32,279	P, R, F
Turnbull et al. (2008)	Auto Tag.	Music	CO	AF, MOD	-	Prediction	500	P, R, F
Cao et al. (2009)	Tag. rec.	Bookmarks	FO	CC, PE	Bibsonomy	Prediction	274,139 378,378	P, R, F
De Meo et al. (2009)	Tag. rec.	Bookmarks	FO	CC	Delicious	User assess.	160	P, R, F
Marinho et al. (2009)	Tag. rec.	Bookmarks	FO	PR, PE	Bibsonomy	Prediction	378,378	P, R, F
Martínez et al. (2009)	Auto Tag.	Sound	CO	AF, SIM	Freesound	User assess.	260	[a]
Rendle & Schmidt-Thieme (2009)	Tag. rec.	Bookmarks	FO	PR, PE	Bibsonomy	Prediction	-	P, R, F
Wu et al. (2009)	Tag. rec.	Image	FO, CO	CC, IF, MOD	Flickr	User assess.	500	Variant of R
Zhang et al. (2009)	Tag. rec.	Bookmarks	FO, CO	CC, SIM	Bibsonomy	Prediction	464,930	P, R, F
Ballan et al. (2010)	Tag. rec.	Video	FO, CO, CT	CC, IF, SIM	YouTube	User assess.	56	P
Chen et al. (2010b)	Tag. rec.	Video	CT	PK	YouTube	-	-	-
Ivanov et al. (2010)	Tag. rec.	Image	CO	IF, SIM	Flickr, Goolge	Prediction	3,200	P, R, F
Lee et al. (2010)	Tag. rec.	Image	FO, CO	IF, SIM, PR	Flickr	Prediction	25,000	Variant of $S@\kappa$
Liu et al. (2010a)	Auto Tag.	Image	FO, CO, CT	CC, IF, SIM	Flickr	User assess.	50,000	P, R, F
Rae et al. (2010)	Tag. rec.	Image	FO	CC, PR	Flickr	Prediction	3,000	MRR, MAP, $P@5$
Sevil et al. (2010)	Tag. rec.	Image	FO, CO	IF, SIM	Flickr	Prediction	15,000	$P@8$, $P@20$, $P@25$
Toderici et al. (2010)	Tag. rec.	Video	CO	IF, VF, AF, MOD	YouTube	User assess.	500	P
Lops et al. (2012)	Tag. rec.	Bookmarks	FO, CO	TF, SIM	Bibsonomy	Prediction	378,378	P, R, F
Sordo (2012)	Auto Tag.	Music	CO	SIM	-	Prediction	8,790	P, R, F

[a] Average over user judgement scores or agreement on judgement scores.
[b] Percentage of taglines that contain at least one recommended tag.

ABBREVIATIONS: FO=folksonomy-based; CO=content-based; CT=context-based sources; IF=image features; AF=audio features; TF=text features; CC=tag co-occurrence in resources; PR=probabilistic model; CF=collaborative filtering; PK=PageRank algorithm; FK=FolkRank algorithm; SIM=resource similarity tag propagation; MOD=model-based tag prediction; GPS=geolocation data.

Table 2.4: Summary of tag recommendation and related auto tagging methods proposed in the literature. Entries in the table are sorted chronologically. For details on evaluation measures see Sec. 2.3.4.

For that, we require of a training set with examples of resources for each tag that has to be modelled. Using these examples, a machine learning algorithm can learn the relations between low-level features and tags. Thus, given a new resource, the algorithm can predict whether a particular tag is potentially relevant. Examples of this approach have been proposed for recommending tags for image resources (Barnard et al., 2003; Li & Wang, 2008; Anderson et al., 2008; Chen et al., 2008; Wu et al., 2009), audio resources (Barrington et al., 2007; Turnbull et al., 2008), and video resources (Toderici et al., 2010).

The other common approach in content-based tag recommendation systems is the propagation of tags from resources that have already been annotated to resources that have not yet been annotated (marked with the abbreviation "SIM" in Table 2.4). In this case, the extracted low-level features define a multi-dimensional feature space in which similarity measures can be defined. Then, given a resource, other similar resources can be retrieved. Hence, in these systems, tags can be propagated among similar resources. Examples of such content-based tag recommendation systems can be found in the image (Liu et al., 2010a; Ivanov et al., 2010; Lee et al., 2010; Sevil et al., 2010), audio (Martínez et al., 2009; Sordo, 2012), video (Ballan et al., 2010), and bookmark domains (Zhang et al., 2009; Lops et al., 2012).

The main advantage of tag recommendation systems based on content analysis is that, after the training step, these can be applied to any resource, even if there is no associated metadata. Conversely, a disadvantage is that these systems are not directly generalisable to other domains because the feature extraction and similarity definition steps are highly dependent on the type of resources for which tags are recommended. Similarly, even when staying in the same domain, content-based systems do not always generalise to resources which are sufficiently different from those used in the training step. For instance, a system that learns to recommend tags for music resources and that is trained with a dataset of electronic music, will not necessarily perform well when tested on a dataset of classical music. Moreover, the content analysis and training steps of content-based systems are, in general, computationally expensive, particularly if a model needs to be built for every potentially suggested tag. Finally, in many cases, the training step of content-based models requires great human effort in building and validating the datasets from which the models will be learnt.

2.3.2 Based on folksonomy analysis

An important number of works on tag recommendation systems are based on analysing the folksonomy of a tagging system in order to provide recommendations given a resource and a number of already existing tags assigned to that resource (i.e., the input tags). As opposed to content-based systems, folksonomy-based systems are easily generalisable to several domains because

the content of the resources is not analysed. However, in order to provide recommendations for a given resource, folksonomy-based systems require the presence of some already assigned tags.

To formally define a folksonomy, the adoption of the tripartite graph model proposed by Mika (2007) is very common. This model is very similar to previous models proposed by Hotho et al. (2006) and Jäschke et al. (2007). In Mika's model, the folksonomy is represented as a tripartite graph in which users, tags and resources are included as nodes, and edges establish ternary relations among them, the so called tag applications (Sec. 1.2). The tripartite graph can be unfolded into a bipartite graph after discarding one of the three sets of nodes (i.e., users, tags or resources). In this way, it is possible to obtain a graph which relates tags and users, a graph which relates users and resources, and a graph which relates tags and resources. This last bipartite graph is the view of the folksonomy that we work with in this thesis, as it allows the derivation of relations between tags on the basis of their shared resources and vice versa. Details on the formal definition of the model can be found in Chapter 3 of this thesis (Sec. 3.2). The graph that relates users and resources is not exploited in this thesis because it does not bring any particular information about tags. Furthermore, the graph that relates tags and users is neither used in this thesis because the tag recommendation systems we propose are not personalised to users' particular tagging styles (see below).

Most of the tag recommendation methods based on folksonomy analysis take advantage of the co-occurrence of tags in resources in order to estimate a similarity or relatedness measure between pairs of tags[22] (these methods are marked with the abbreviation "CC" in the "based on" column of Table 2.4). The general assumption is that tags that tend to appear together in the taglines of annotated resources are potentially similar. Using the bipartite graph relating tags and resources, it is thus possible to define a tag-tag similarity matrix in which every element indicates the number of resources in which two tags co-occur (i.e., the number of resources in which two tags appear together). By applying a normalisation to this matrix, several similarity measures can be obtained, such as the widely adopted cosine similarity or the Jaccard index (Mika, 2007; Markines et al., 2009). For more details on the construction of this similarity matrix and the normalisation process we again refer the reader to Chapter 3 of this thesis (Sec. 3.2.1).

Given a tag-tag similarity matrix, it is possible to retrieve a ranked list of the most similar tags to a given target tag. This is the basis of many folksonomy-based tag recommendation systems. Considering a set of input tags, potentially

[22]The words "similar" and "related" have, in fact, different meanings. Two tags can be considered similar if there is a considerable overlap in their meanings, or can be considered related if they convey complementary information. When talking about tags, in this thesis we use the term "similar" in a broad sense, referring to any sort of potentially relevant relation between tags.

relevant tags can be obtained by retrieving the most similar tags to each input tag. Then, these tags are aggregated into a single ranked list of candidate tags by taking into account the similarity scores with their respective input tag (Sigurbjörnsson & Zwol, 2008). In general, the top tags of the aggregated set of candidates are selected to generate the output of the tag recommendation system. Tag recommendation systems based on this approach have been proposed for the image (Anderson et al., 2008; Sigurbjörnsson & Zwol, 2008; Garg & Weber, 2008; Wu et al., 2009; Liu et al., 2010a; Rae et al., 2010), video (Ballan et al., 2010), and bookmark domains (Lipczak, 2008; Song et al., 2008; Cao et al., 2009; De Meo et al., 2009; Zhang et al., 2009). Both aggregation and selection procedures are applicable not only to folksonomy-based tag recommendation systems but also to those content-based systems based on tag propagation from similar resources.

As an alternative to the folksonomy-based recommendation methods that use tag co-occurrence information, some authors have also proposed focusing on resource similarity in order to generate sets of candidate tags (again marked with the abbreviation "SIM" in Table 2.4). In this case, resource similarity can be defined on the basis of the tags shared among resources (i.e., the more tags two resources share, the more similar they potentially are). Sevil et al. (2010) and Lops et al. (2012) propose an approach that, given a number of input tags or keywords extracted from a textual description of a resource, it queries a folksonomy to retrieve the tags of other resources whose taglines include at least one of the input tags or keywords. These tags can then be used as candidates. An alternative approach also based on the concept of resource similarity is the one proposed by Lee et al. (2010). In this approach, given an image with some assigned tags, a subset of the main folksonomy is built by considering information from all tag applications that involve any image resource whose tagline includes at least one of the input tags. Then, given this subset of the folksonomy, a probabilistic method is used to compute the probability of each tag being relevant for the new resource.

Besides considering the relations between tags and resources, some works on folksonomy-based tag recommendation put some emphasis on the relations between tags and users. These systems focus on the personalisation of the recommendation process, and are particularly suited to promote and reinforce specific user's tagging conventions (marked with the abbreviation "PE" in Table 2.4). Examples of personalised tag recommendation systems include the approaches proposed by Jäschke et al. (2007), based on collaborative filtering techniques and on FolkRank, an adaptation of the PageRank algorithm (Brin & Page, 1998). Also, Garg & Weber (2008) and Lipczak (2008) propose methods in which the vocabulary of tags previously employed by the user annotating a resource is particularly promoted during the recommendation process. Furthermore, other approaches take advantage of probabilistic and machine learning techniques to learn latent interactions between users, resources and

tags, instead of only resources and tags (Rendle & Schmidt-Thieme, 2009; Marinho et al., 2009), or derive tag-tag similarity matrices not only on the basis of shared resources but also on the basis of shared users (Cao et al., 2009).

Some authors also propose tag recommendation methods that combine folksonomy-based and content-based approaches. These methods typically generate separate lists of candidate tags using any of the techniques described above and in the previous section. Then, the lists of candidates are aggregated to create a single set of tags that can be recommended. Intuitively, the combination of folksonomy-based and content-based approaches allows the design of tag recommendation systems featuring the advantages of both approaches, but these systems also become more complex. To the best of our knowledge, no formal comparison of these approaches has been carried out in the tagging literature. Methods combining both approaches have been proposed for the image (Anderson et al., 2008; Wu et al., 2009; Liu et al., 2010a; Lee et al., 2010; Sevil et al., 2010), video (Ballan et al., 2010), and bookmark domains (Zhang et al., 2009; Lops et al., 2012).

As we can see, a wide variety of methods have been proposed for the task of tag recommendation. However, it is worth mentioning that the output of most of the previously mentioned methods consists, in fact, of a ranked list of candidates from which the top tags are recommended. Very few articles give further explanations on how that number of top tags can be chosen to optimise the relevance of the final set of recommended tags. Those few articles propose rather simple heuristics like systematically recommending half of the candidate tags (Cao et al., 2009), recommending as many tags as users employ on average (Marinho et al., 2009; Rendle & Schmidt-Thieme, 2009), or recommending those tags that are most times repeated in the lists of candidates before aggregation (Martínez et al., 2009; Sordo, 2012). Sood et al. (2007) extends this and proposes the recommendation of candidate tags whose score is above the mean score of all candidates. In general, existing literature does not put emphasis on the selection of the number of tags to recommend. To counteract the lack of methods for solving this problem, in this thesis we propose several strategies with which the selection of tags to recommend can be approached (Chapter 3).

Furthermore, we see from the literature review that some tag recommendation methods are focused on the personalisation of recommendations to user's particular tagging conventions. In general, these systems are most useful in the context of broad folksonomies where users can be encouraged to tag for self-organisation purposes and, therefore, the reinforcement of particular user's tagging conventions is not necessarily a problem but a desirable feature (see Secs. 2.2.2 and 2.2.3). However, in tagging systems featuring narrow folksonomies, a common tagging style across users is preferred so that resources are tagged more uniformly and with a common vocabulary (Lipczak, 2008). In this thesis, we propose a novel approach for tag recommendation in which tag

suggestions are *personalised* to groups of resources instead of users. In this way, we aim to leverage tag and resource relations which are particular to specific classes of resources, and therefore reinforce a common vocabulary for each class of resources (Chapter 4).

2.3.3 Based on contextual data

Besides content-based and folksonomy-based tag recommendation, there has been some work on using contextual data retrieved from external sources (marked with the abbreviation "CT" in the "type of approach" column of Table 2.4). In general, external data is retrieved in order to complement what can be extracted from the folksonomy or from the analysis of resources' content. However, in some cases, contextual data is the only source of information used to generate the recommendation.

Sood et al. (2007) propose a tag recommendation system for blog posts which finds posts with similar textual content by querying a blog aggregator service. Then, tags already present in these similar posts are suggested for the target post. This is the same idea we have seen for content-based and folksonomy-based systems that propagate tags from similar resources (those marked with the abbreviation "SIM" in Table 2.4). The difference in this case is that the actual similar resources come from external sites instead of the same tagging system or online sharing platform. In a similar vein, Chen et al. (2010b) propose a system for video tagging in which, given the title and some assigned tags, a search engine finds written textual content about that video hosted on blogs or other online sites. Then, the textual content is processed, and a number of keywords are extracted to be suggested as tags. That approach requires however that the resource being annotated is known enough so that people have written about it (e.g., a well known movie). Hence, a priori, it is not suitable for user generated content. A completely different idea is proposed by Naaman & Nair (2008), which describe a tag recommendation system for a mobile application that allows users to take pictures and upload them to Flickr. Using the geolocation data provided by the GPS signal, the recommendation system is able to query Flickr for photos taken in the same place, and use the tags of these photos as candidates for recommendation. Finally, in the methods proposed by Liu et al. (2010a) and Ballan et al. (2010), an external lexical database is used. Liu et al. (2010a) propose to expand the set of candidate tags by adding synonyms and hypernyms of the already present tags retrieved from WordNet (Miller, 1995). Similarly, Ballan et al. (2010) propose to use WordNet in order to expand the set of tags already present to a given resource being annotated, and to be able in this way to perform a broader tag-based resource similarity search returning more candidate tags.

2.3.4 Evaluation of tag recommendation strategies

Given a set of recommended tags for a particular resource, deciding which tags are relevant and which are not is a highly subjective and difficult task. In general, evaluation strategies are either based on a standard information retrieval prediction task or are based on user assessment of the generated recommendations (Table 2.4). Also, with the exception of the Bibsonomy datasets prepared for the 2008 and 2009 ECML PKDD Discovery Challenges[23], there are no well established datasets for tag recommendation systems (Wang et al., 2012). In general, Flickr data is used to evaluate tag recommendation systems dealing with image content, and Bibsonomy data is used in the case of recommendation systems targeted at bookmarks.

In prediction-based evaluation, it is standard practice to use a number of annotated resources as ground truth, which is further divided into a training set and a testing set. Tag recommendation systems are trained with all annotated resources in the training set, while the testing set is used to evaluate the ability of the recommendation system for predicting the original taglines of the resources. Because many recommendation systems require the presence of some input tags in order to provide recommendations, the taglines of the resources in the testing set are typically divided into a set of input tags, and the set of tags that has to be predicted. In essence, some tags are removed from the taglines of the resources in the testing set. Thus, the goal of the recommendation system is the prediction of the removed tags given the remaining input tags. In information retrieval terms, the removed tags become the *relevant* tags that the system has to predict.

This kind of evaluation approach is very useful, as it allows a systematic assessment of recommendation systems that can be tested under different parameter settings, system configurations, and with a huge number of evaluated resources. However, in most cases, the datasets are formed by user provided resources and the annotations are gathered from an online sharing system. Thus, these are not curated or assessed by experts. For example, annotations may include wrongly assigned or redundant tags. Hence, both the training and the testing sets contain potentially noisy data (Doerfel & Jäschke, 2013). Moreover, a tag suggested by a recommender system to a given resource is only considered correct if that given resource was originally annotated with that tag (i.e., the tag is part of the set of removed tags). It is a well-known problem of folksonomies that descriptions tend to be scarce and not coherent across resources (Sec. 2.2.4). Thus, it can easily happen that recommended tags that are actually relevant for a given resource are not part of the set of removed tags and are not considered correct. Furthermore, the recommendation system might

[23] European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (`http://www.kde.cs.uni-kassel.de/ws/rsdc08/` and `http://www.kde.cs.uni-kassel.de/ws/dc09/`).

recommend variations of tags that are in the list of removed tags (e.g., plural forms or synonyms), but that are considered incorrect as are not exactly the same. For these reasons, the results provided by these kind of evaluations are considered to be an underestimate of the real performance of the system (Garg & Weber, 2008). In essence, the real performance of the system can not be accurately assessed because there is not a unique and single solution for a tagging task. Thus, the overall performance of a tag recommendation system highly depends on the evaluation criteria, and no "gold standards" can be defined as the maximum performance for a given system.

Following standard practice in information retrieval, prediction-based evaluation approaches tend to use precision (P), recall (R), and f-measure (F) metrics averaged over all evaluated resources (Manning et al., 2008). In the tag recommendation context, those measures are defined as follows:

$$P = \frac{|\mathbf{T}_\mathrm{R} \cap \mathbf{T}_\mathrm{D}|}{|\mathbf{T}_\mathrm{R}|} \ , \ R = \frac{|\mathbf{T}_\mathrm{R} \cap \mathbf{T}_\mathrm{D}|}{|\mathbf{T}_\mathrm{D}|} \ , \text{ and } F = \frac{2PR}{P+R}, \tag{2.1}$$

where $\mathbf{T}_\mathrm{R}$ is the set of tags recommended by the system and $\mathbf{T}_\mathrm{D}$ is the set of tags removed from a resource for evaluation purposes (i.e., the set of relevant tags). As most of the works on tag recommendation do not limit the number of tags to recommend, these measures are normally applied for different values of κ recommended tags, that is to say, only considering the top κ recommendations outputted by the system (i.e., $P@\kappa$, $R@\kappa$, $F@\kappa$).

P, R and F are the most common measures, but other measures have been also used (Table 2.4). The most important ones include the Success at rank κ, $S@\kappa$, the Mean Reciprocal Rank, MRR, and the Mean Average Precision, MAP (Manning et al., 2008). $S@\kappa$ is a relaxed version of P which indicates the presence or absence of at least one relevant tag in the first κ recommended tags. MRR computes the inverse of the position where the first relevant tag appears in a set of recommended tags. Finally, MAP is a measure designed to particularly take into account the order in which the recommended tags are outputted by the system.

Besides evaluation strategies based on a tag prediction task, some works also evaluate tag recommendation systems through user assessment. User-based evaluation does not allow the systematic evaluation of tagging systems, but gives a point of view which is, a priori, closer to a real-world evaluation of the system. In user-based assessment, the set of relevant tags for a given resource is not defined by tags removed from a resource ($\mathbf{T}_\mathrm{D}$), but by user judgement over the appropriateness of recommended tags. Moreover, user-based evaluation allows the collection of qualitative user feedback that can shed some light on relevant aspects of the recommendation process. Therefore, user-based evaluation and prediction-based evaluation can be complementary strategies.

The few user-based evaluations found in the tag recommendation literature typically consist of the validation of the appropriateness of each recommended

tag for a given resource. Thus, users provide a judgement of how relevant each recommended tag is for the resource at hand, typically over an n-point scale (see e.g., Sigurbjörnsson & Zwol, 2008; De Meo et al., 2009). If more than one user judgement is performed for a particular tag application, these can be averaged and then binarised to determine whether a particular tag is relevant or not for a given resource. In this way, a set of relevant tags among the recommended tags can be determined, and the previously described evaluation measures can be applied to perform user-based evaluation.

Another approach for user-based evaluation consists of the use of a prototype with which users annotate resources. This evaluation mimics a real-world situation in which users annotate resources with the help of a tag recommendation system. Using this approach, user activity can be logged and collected for further analysis, and users implicitly select relevant tags from the list of recommendations by adding them to the tagline of the resource at hand. To the best of our knowledge, only two works of those found in the literature perform such kind of evaluation. Jäschke et al. (2009) perform a small evaluation based on a real-world scenario where users have to tag bookmarks in Bibsonomy. Specifically, P and R metrics are computed by comparing tag recommendations performed to every bookmark and the final taglines that users introduced. Similarly, Naaman & Nair (2008) use a prototype to evaluate tag recommendations in an image tagging mobile application. In essence, these kind of evaluations allow to identify which of the tags recommended during the annotation process are finally "accepted" by users in a real-world scenario. Furthermore, additional information can be obtained through qualitative feedback provided by users and through the analysis of interaction patterns, which can give insights on aspects such as the time required for users to annotate resources or the difficulty of the annotation process. Noticeably, we are not aware of any user evaluation based on prototypes performed in the context of a large-scale and real-world tagging system. In this thesis, we follow both prediction-based and user-based evaluation evaluation methodologies, using prototype and tag assessment strategies, in order to comprehensively assess the successfulness and impact of the tag recommendation systems we describe.

2.3.5 Impact of tag recommendation

Once a tagging system has a critical mass of users, its underlying folksonomy is supposed to reach a point in which the vocabulary and tagging patterns are mature enough to allow proper indexing, browsing and searching of the content (see Sec. 1.2 and Guy & Tonkin, 2006; Spiteri, 2007). Additionally, it leverages the value of the folksonomy as a source of knowledge mining (Wagner et al., 2014). In the tagging literature, vocabulary maturity is understood as the point of consensus in which a certain set of tags and tagging conventions are widely adopted by most users of the system (Halpin et al., 2006; Sen et al., 2006; Cattuto, 2006; Sood et al., 2007; Robu et al., 2009; Wagner et al., 2014).

The emergence of consensus depends on several factors. Halpin et al. (2006) and Robu et al. (2009) state that consensus emerges as a combination of the background knowledge that is shared by users of a tagging system, and by the way in which users annotating resources are exposed to the annotations performed by other users. Similarly, Sen et al. (2006) suggest that users' choice of tags is influenced by their personal beliefs (background knowledge) and the tagging conventions of the community. The "social proof" theory supports that idea, as it states that users tend to consider as correct those annotation conventions that other users have already employed (Cialdini, 2003). Therefore, the more users are exposed to the tagging conventions of other users, the faster the consensus should emerge. Suggesting potentially relevant tags at annotation time can greatly contribute to user's exposure to other tagging conventions. Hence, it has been argued that tag recommendation systems can have a big impact on the convergence to consensus in a folksonomy. As stated by Marlow et al. (2006), "a suggestive system may help consolidate the tag usage for a resource, or in the system, much faster than a blind tagging system would. A convergent folksonomy is more likely to be generated when tagging is not blind". This idea is also highlighted by other authors (Golder & Huberman, 2006; Jäschke et al., 2007; Sood et al., 2007; Farooq et al., 2007; Robu et al., 2009; Wagner et al., 2014).

Some studies have been focused on measuring the vocabulary convergence in a folksonomy. A common way in which this aspect is measured is through the analysis of the distribution of tags' frequency of occurrence in a folksonomy. In studies analysing natural language, it has been observed that the distribution of the frequency of occurrence of words tends to follow a power law distribution (Solé, 2005; Cattuto, 2006). Hence, some authors suggest that folksonomies whose distribution of tag frequency can be fitted by a power law, exhibit mature vocabularies (Mathes, 2004; Cattuto, 2006; Halpin et al., 2006; Wagner et al., 2014). In these kind of studies, the word "consensus" is typically used to indicate agreement on how different users annotate a particular resource. Therefore, those are targeted to tagging systems featuring broad folksonomies, in which a single resource can be annotated several times by distinct users. Several empirical studies have observed the emergence of power law distributions in the folksonomies of different tagging systems, not only in broad folksonomies like Delicious, but also in narrow folksonomies like Flickr (Cattuto, 2006; Halpin et al., 2006; Guy & Tonkin, 2006; Sigurbjörnsson & Zwol, 2008; Robu et al., 2009; Wagner et al., 2014).

Another way to look at folksonomy vocabulary consensus is by analysing tagging behaviour and the way in which tags are shared across users. In a study by Farooq et al. (2007), the folksonomy of CiteULike is analysed, and a number of basic metrics are proposed to quantify some of its characteristics. In particular, authors observe that the rate at which new tags are created maintains a high correlation with the rate at which new users start using the tagging system.

This suggests that tags are not much shared among users, and that users tend to develop their own personal tag vocabularies. Furthermore, Marlow et al. (2006) analysed data from Flickr's folksonomy focusing on how users belonging to different groups share tags in their vocabulary[24]. The results indicated that pairs of users belonging to a same group are much more likely to share tags than pairs of random users. That implies a stronger influence among users of the same group, probably because their shared cultural background is stronger and because they are more implicitly exposed to the tagging conventions of other users in the same group. As a general conclusion, both studies suggest that tag recommendation could greatly contribute in increasing and consolidating the vocabulary sharing among users of a tagging system. The same idea is shared by Golder & Huberman (2006), Jäschke et al. (2007), and Sood et al. (2007). According to Sood et al. (2007), by using a tag recommendation system, users can see how other users tag the resources and can then better choose when to reuse already existing tags or when to create new ones, which contributes to the stabilization of the vocabulary. Similarly, Zangerle et al. (2011) perform a study on *hashtag* recommendation for Twitter[25], a microblogging site, and hypothesise that the use of hashtag recommendation should help homogenising hashtags.

Besides the impact expected in the vocabulary sharing and folksonomy consensus, some authors also suggest other problems of tagging systems that can be lessened by using a tag recommendation system. Naaman & Nair (2008) performed an empirical study comparing two mobile phone applications for tagging and uploading photos to Flickr where one of them featured a tag recommendation system. Authors observed that the taglines of photos uploaded using the application with the tag recommendation system had, on average, more tags than the taglines of the photos uploaded with the other system. Therefore, tagging systems can also contribute to mitigate the problem of tag scarcity or lack of tags. Similarly, Jäschke et al. (2007; 2012) hypothesise that tag recommendation simplifies the process of finding good tags for the resources being described and thus increases the chances of getting resources annotated. Also, Wang et al. (2012) hypothesise that tag recommendation can improve both the quality of tags and the efficiency of the tagging process, by clarifying the semantics of tags and reducing the manual cost of tagging. Finally, Sood et al. (2007) suggest that a tag recommendation system fundamentally changes the tagging process from being a generation process, where users must create tags from scratch, to being a recognition process, where users have to recognise valid tags from a list of suggestions. As a result, it can also help alleviate typical synonymy problems by suggesting specific variants of tags.

As it can be seen, there has been considerable discussion in the literature

[24] In Flickr, users can explicitly be members of groups that, for example, share an interest for particular kinds of photos (Table 2.1).

[25] `http://www.twitter.com`

regarding the expected impact of tag recommendation systems in the folksonomies of tagging systems. However, we are not aware of any comprehensive study performing a deep analysis of the impact of a tag recommendation system into a real-world and large-scale folksonomy. To our knowledge, this thesis includes, in Chapter 5, the first empirical analysis of this kind.

CHAPTER 3

A scheme for folksonomy-based tag recommendation

3.1 Introduction

In this chapter we describe a general scheme for tag recommendation in large-scale tagging systems. The approach we describe here is based on tag co-occurrence in folksonomies, meaning that we do not perform any content analysis of the information resources for which we produce tag recommendations. We uniquely rely on the tag co-occurrence information that can be derived from the folksonomy itself. As the scheme we describe only relies on this information, it is rather domain-independent and could be easily adapted to other tagging systems, either alone or as a complement of perhaps more specific content-based strategies. Hence, our approach is highly related to those outlined in Sec. 2.3.2 and, in particular, to the tag recommendation methods described by Sigurbjörnsson & Zwol (2008) and Garg & Weber (2008).

A particularly interesting aspect of our tag recommendation scheme, which differentiates it from previous works, is a step focused on automatically selecting the number of tags to recommend. That step is accomplished by considering the relative relevance scores of a set of candidate tags with respect to a set of input tags (see below). Other tag recommendation methods generally do not consider this aspect and evaluate their solutions at different numbers of κ recommended tags (Sec. 2.3.2). This is an unrealistic situation as, in a real-world scenario, only a limited number of tags can be suggested to users, and an arbitrary decision of this number may yield suboptimal recommendations either missing relevant tags or suggesting too many non-relevant ones.

We propose eight tag recommendation methods which are based on the aforementioned general scheme. The proposed methods, jointly with several baselines,

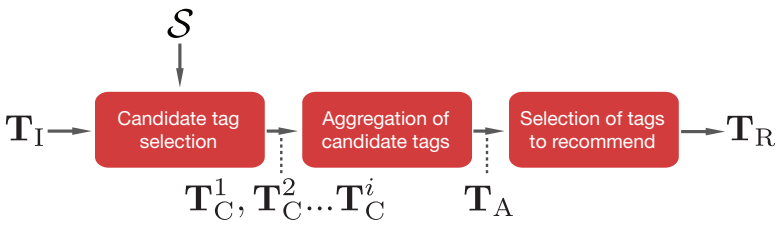


Figure 3.1: Block diagram of the described tag recommendation scheme.

are evaluated with data coming from Freesound and Flickr. For the best scoring methods, we also analyse the influence of their configurable parameters. Overall, we perform more than 100 different experiments and compute around seven million tag recommendations.

The rest of this chapter is organized as follows. In Sec. 3.2 we describe the different steps of our tag recommendation scheme and the strategies we propose to compute each step. Then, in Sec. 3.3, we outline the characteristics of the evaluation datasets and describe the methodology we followed to evaluate our methods and the considered baselines. The results of our evaluation are reported in Sec. 3.4, and the chapter ends with a discussion about our findings and future work (Sec. 3.5).

3.2 Method

Given a set of input tags $\mathbf{T}_\mathrm{I}$ and a tag-tag similarity matrix $\mathcal{S}$ derived from a folksonomy $\mathcal{F}$, the general scheme for tag recommendation outputs a set of recommended tags $\mathbf{T}_\mathrm{R}$ (Fig. 3.1). The described scheme is composed of three independent steps: 1) Candidate tag selection, 2) Aggregation of candidate tags, and 3) Selection of 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; De 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: Candidate tag selection. Given $\mathbf{T}_\mathrm{I}$ and a tag-tag similarity matrix $\mathcal{S}$ derived from $\mathcal{F}$, this step retrieves a set of θ candidate tags $\mathbf{T}_\mathrm{C}^i$ for each input tag $\mathbf{T}_{\mathrm{I}_i}$.
- Step 2: Aggregation of candidate tags. This step takes the sets of candidates $\mathbf{T}_\mathrm{C}^i$, assigns a score value to each individual tag, and aggregates all candidates to form a single list of tags with assigned scores $\mathbf{T}_\mathrm{A}$.

- Step 3: Selection of tags to recommend. This step automatically selects which tags to recommend given the candidate tags and score values of $\mathbf{T}_{\mathrm{A}}$. The output of this step is the final set of recommended tags $\mathbf{T}_{\mathrm{R}}$.

3.2.1 Candidate tag selection

We start the recommendation process by obtaining a number of related candidate tags to the set of input tags $\mathbf{T}_{\mathrm{I}}$. For each input tag $\mathbf{T}_{\mathrm{I}_i}$, we get a set of candidates $\mathbf{T}^i_{\mathrm{C}}$ by selecting the θ closest tags to $\mathbf{T}_{\mathrm{I}_i}$ according to a tag-tag similarity measure. For this purpose, we build a tag-tag similarity matrix $\mathcal{S}$ based on the tag assignment information contained in the folksonomy $\mathcal{F}$. Note that $\mathcal{S}$ is not dependent of the particular $\mathbf{T}_{\mathrm{I}_i}$ for which we are selecting candidates. Therefore, it only needs to be computed once for a given $\mathcal{F}$[26].

To represent the folksonomy $\mathcal{F}$, we use the model proposed by Mika (2007). Mika's model considers three finite sets of objects $\mathbf{A}$, $\mathbf{C}$ and $\mathbf{I}$, which correspond to "actors" (i.e., users), "concepts" (i.e., tags) and "instances" (i.e., resources), respectively. In this thesis, instead of the variables $\mathbf{A}$, $\mathbf{C}$ and $\mathbf{I}$ defined by Mika (2007), we employ the notation $\mathbf{U}$, $\mathbf{T}$ and $\mathbf{R}$, which more closely relates to the "users", "tags" and "resources" terminology that we use. The sets of users, tags and resources are represented as nodes in the graph such that the set of nodes $\mathbf{V}$ is defined as $\mathbf{V} = \mathbf{U} \cup \mathbf{T} \cup \mathbf{R}$. The ternary relations between a user, a tag and a resource (i.e., tag applications) are then represented as the edges of the graph $\mathbf{E} = \{\{u, t, r\} | (u, t, r) \in \mathcal{F}\}$. Hence, the graph $\mathcal{G}$ that represents a folksonomy $\mathcal{F}$ is finally defined as $\mathcal{G}(\mathcal{F}) = \langle \mathbf{V}, \mathbf{E} \rangle$.

We unfold $\mathcal{G}(\mathcal{F})$ into the bipartite graph $\mathcal{TR}$, which only reflects the associations between tags and resources. The bipartite graph $\mathcal{TR}$ can be represented 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 $\mathcal{S}$ so that

$$\mathcal{S} = \mathcal{D}\mathcal{D}', \tag{3.1}$$

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 labelled 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 the aforementioned similarity measures, we apply different

[26] As is described later in Sec. 3.3.1, we filter out the least frequent tags of our folksonomy in order to reduce the computational complexity of $\mathcal{S}$.

normalisation procedures to $\mathcal{S}$. Cosine similarity is defined as

$$s_{t_i,t_j} = \frac{\sum_n d_{i,n} d_{j,n}}{\sqrt{\sum_n {d_{i,n}}^2}\sqrt{\sum_n {d_{j,n}}^2}}. \tag{3.2}$$

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). Therefore, cosine similarity can be obtained by dividing each element in $\mathcal{S}$ (Eq. 3.1) by $\sqrt{s_{t_i,t_i}}\sqrt{s_{t_j,t_j}}$. Similarly, the Jaccard index is defined as

$$s_{t_i,t_j} = \frac{\sum_n d_{i,n} d_{j,n}}{\sum_n {d_{i,n}}^2 + \sum_n {d_{j,n}}^2 - \sum_n d_{i,n} d_{j,n}}, \tag{3.3}$$

which is equivalent to dividing each element in $\mathcal{S}$ by $s_{t_i,t_i} + s_{t_j,t_j} - s_{t_i,t_j}$. Independently of the similarity measure, $\mathcal{S}$ can be represented as a graph where nodes correspond to tags and edges represent the similarities between two tags (Fig. 3.2).

Once we have a tag similarity matrix $\mathcal{S}$, we iterate over the input tags $\mathbf{T}_\mathrm{I}$ and get, for each element $\mathbf{T}_{\mathrm{I}_i}$, a set of candidates $\mathbf{T}^i_\mathrm{C}$. Specifically, we select the θ most similar tags to $\mathbf{T}_{\mathrm{I}_i}$ (i.e., the θ most similar graph neighbours of $\mathbf{T}_{\mathrm{I}_i}$) 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 3θ candidate tags (separated into three sets), provided that all three input tags have at least θ graph neighbours.

3.2.2 Aggregation of candidate tags

The next step of our tag recommendation scheme takes all the sets of candidates $\mathbf{T}^i_\mathrm{C}$, assigns a score value ϕ_j to every candidate $\mathbf{T}^i_{\mathrm{C}_j}$ in $\mathbf{T}^i_\mathrm{C}$, and then aggregates all sets into a single list of tags with assigned scores $\mathbf{T}_\mathrm{A}$. The output of this step, $\mathbf{T}_\mathrm{A}$, is a list of tuples where each element contains a tag and an assigned score. To accomplish this step, we propose two different strategies:

Similarity-based Strategy

In the Similarity-based Strategy, the j-th candidate tag $\mathbf{T}^i_{\mathrm{C}_j}$ of $\mathbf{T}^i_\mathrm{C}$ is assigned a score ϕ_j that directly corresponds to the similarity value between the candidate tag and the corresponding input tag $\mathbf{T}_{\mathrm{I}_i}$, i.e., $\phi_j = s_{x,y}$, where $x = \mathbf{T}^i_{\mathrm{C}_j}$ and $y = \mathbf{T}_{\mathrm{I}_i}$. After that, the list of tuples $\mathbf{T}_\mathrm{A}$ is constructed as the union of all sets of candidates $\mathbf{T}^i_\mathrm{C}$ and their scores. If a particular tag has duplicates in $\mathbf{T}_\mathrm{A}$ (which can happen if a given tag appears in several sets of candidates $\mathbf{T}^i_\mathrm{C}$), 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

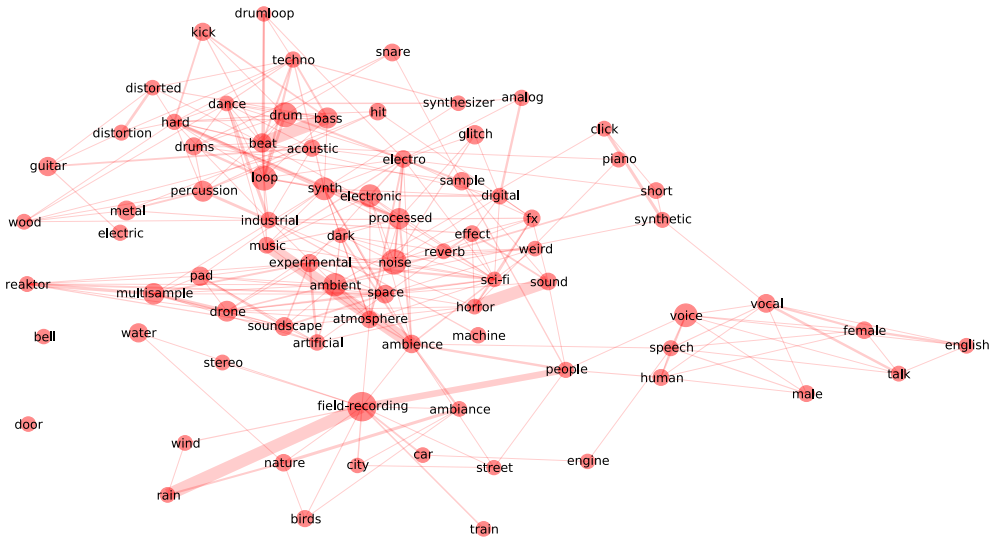


Figure 3.2: Graph visualisation of a tag-tag similarity matrix $\mathcal{S}$ built using cosine similarity and a subset of the Freesound folksonomy. Edge widths represent the cosine similarity between two tags. Tag size is a logarithmic function of the absolute tag frequency. For visualisation purposes, only edges above a certain degree of similarity and tags above a certain level of absolute frequency are shown.

to be similar to more than one input tag. Moreover, as we do not want to recommend tags that are already part of $\mathbf{T}_\mathrm{I}$, we remove any occurrences of these tags in $\mathbf{T}_\mathrm{A}$. We finally normalise the assigned scores by dividing them by the number of input tags $|\mathbf{T}_\mathrm{I}|$.

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 $\mathbf{T}^i_\mathrm{C}$ by similarity values in descending order, and assign scores as $\phi_j = \theta - (n-1)$, where n is the position of the j-th tag in $\mathbf{T}^i_\mathrm{C}$ after sorting (thus n ranges from 1 to θ). Notice that the most similar tag to every input tag will be assigned a score of θ. Even if a particular set $\mathbf{T}^i_\mathrm{C}$ contains less than θ tags (meaning that corresponding input tag $\mathbf{T}_{\mathrm{I}_i}$ has less than θ neighbours in the graph representation of $\mathcal{S}$), the score we assign to the most similar tag will be θ. After score assignment, we proceed exactly as with Similarity-based aggregation: constructing $\mathbf{T}_\mathrm{A}$ as the union of all sets $\mathbf{T}^i_\mathrm{C}$, merging duplicate tags in $\mathbf{T}_\mathrm{A}$ by adding their scores, removing tags appearing in $\mathbf{T}_\mathrm{I}$, and normalising score values by $|\mathbf{T}_\mathrm{I}|$. An example comparing the result of the two aggregation strategies is shown in Table 3.1.

	Aggregated candidate tags ($\mathbf{T}_\mathrm{A}$)			
	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		

Table 3.1: Example of the output of the aggregation step using the Freesound folksonomy with $\mathbf{T}_\mathrm{I} = \{$`field-recording`, `nature`$\}$ and $\theta = 100$. Candidate tags are sorted by their score values. The 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.

3.2.3 Selection of tags to recommend

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

Percentage Strategy

This is a straightforward strategy where ε is determined as a percentage of the highest score in $\mathbf{T}_\mathrm{A}$ by

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

where α is a percentage parameter that must be configured. Following the example shown in Table 3.1, and taking $\alpha = 0.05$, only one tag would be recommended for the Similarity-based aggregation ($\varepsilon = (1-0.05)\cdot 0.307 = 0.292$; $\mathbf{T}_{\mathrm{R}} = \{\texttt{birds}\}$) and three tags would be recommended for the Rank-based aggregation ($\varepsilon = (1-0.05)\cdot 100 = 95$; $\mathbf{T}_{\mathrm{R}} = \{\texttt{birds}, \texttt{ambiance}, \texttt{south-spain}\}$).

Kernel Percentage Strategy

The Kernel Percentage Strategy has two steps. First, we estimate the probability density function PDF of ϕ, the scores of $\mathbf{T}_{\mathrm{A}}$. For that purpose, we use a kernel density estimator (Silvermann, 1986), a fundamental data smoothing technique. The bandwidth of the kernel is automatically determined using Scott's Rule (Scott, 2009). Then, the threshold is defined as the ε that satisfies

$$\int_{\min(\phi)}^{\varepsilon} \mathrm{PDF}(\phi)\,\mathrm{d}\phi = (1-\beta)\int_{\min(\phi)}^{\max(\phi)} \mathrm{PDF}(\phi)\,\mathrm{d}\phi, \tag{3.4}$$

where β is a percentage parameter that must be configured. Therefore, β determines the percentage of the area of the PDF which we consider to include suitable tags for the recommendation (Fig. 3.3). 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 $\mathbf{T}_{\mathrm{A}}$ as a sample extracted from a population of scores with an underlying distribution, the threshold ε can be better determined by considering a percentage of the area of that underlying distribution rather than the percentage of the maximum observed score (as we propose in the Percentage Strategy above).

Statistical Test Strategy

Similarly to the previous strategy, here we also estimate the probability density function PDF of ϕ using a kernel density estimator. However, to determine the threshold ε, we follow an iterative process where, in each iteration, we select a slice of the PDF and perform a statistical test for normality according to

$$\mathrm{AD}(\mathrm{PDF}_{\varepsilon:\max(\phi)}), \tag{3.5}$$

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

$$\varepsilon = \max(\phi) - i \cdot \frac{\max(\phi) - \min(\phi)}{100}, \tag{3.6}$$

where i is the number of the current iteration ($i \in 1, 2, 3, ..., 100$). We stop the iterative process when the test fails for the first time (i.e., when the probability

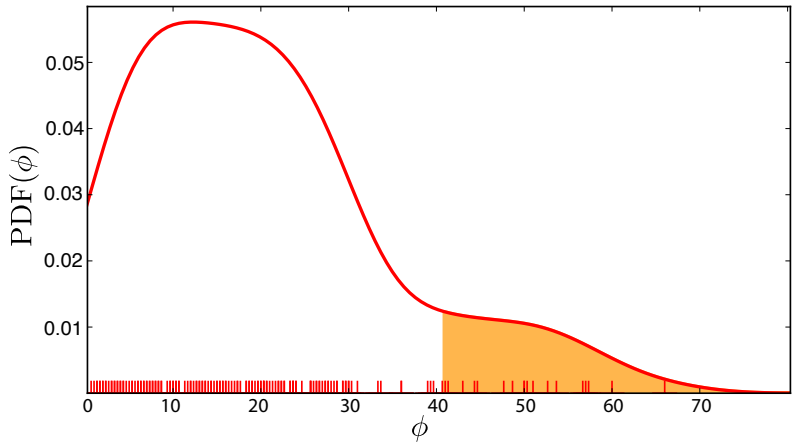


Figure 3.3: Example of the Kernel Percentage Strategy for selecting which tags to recommend (using $\beta = 0.05$). The curve represents the estimated PDF of the scores of $\mathbf{T}_{\mathrm{A}}$. Vertical markers on the horizontal axis show the actual positions of candidate tag scores. The shaded zone in the right of the figure corresponds to the 5% of the total area of PDF. Recommended tags are those under that zone.

of having an independent normal distribution is not statistically significant). The final threshold takes the value of ε at that iteration (Fig. 3.4).

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 and 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 the PDF. 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 normal distribution. We here follow common practice and take this significance level at 0.01 (Scholz & Stephens, 1987). Using another common statistical significance level such as 0.05 would result in less restrictive statistical tests yielding bigger sets of recommended tags.

Linear Regression Strategy

The last strategy we propose consists in calculating the least-squares linear regression of the histogram HIST of ϕ. The threshold is set at the point where the linear regression crosses the vertical axis. The idea behind the Linear Regression Strategy is that, for a given HIST(ϕ), 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 distinguish between both groups at the

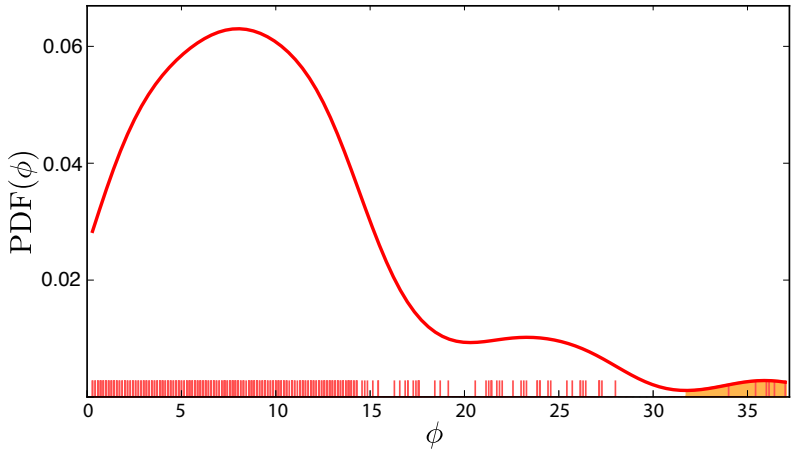


Figure 3.4: Example of the Statistical Test Strategy for selecting which tags to recommend. The curve represents the estimated PDF of the scores of $\mathbf{T}_{\mathrm{A}}$. Vertical markers on the horizontal axis show the actual positions of candidate tag scores. Recommended tags are those under the 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. 3.5 to fail.

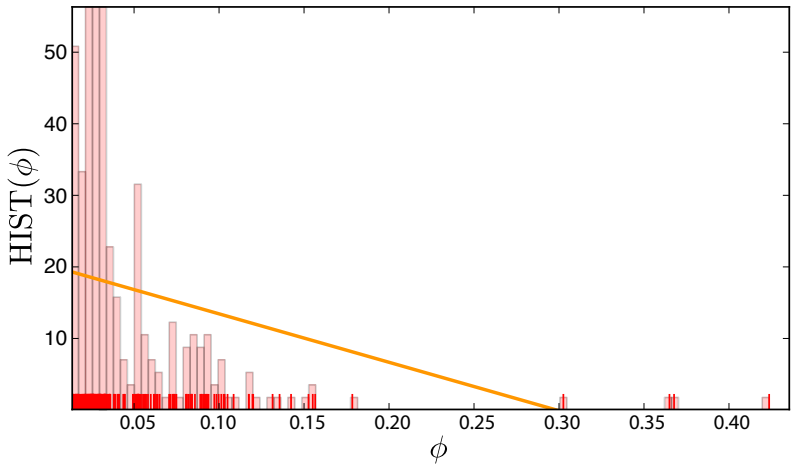


Figure 3.5: Example of the Linear Regression Strategy for selecting which tags to recommend. The straight line shows the linear regression of the histogram HIST of the scores of $\mathbf{T}_{\mathrm{A}}$. Vertical markers on the horizontal axis show the actual positions of candidate tag scores. In this example, the obtained threshold is $\varepsilon \approx 0.29$, which is the point where the linear regression crosses the vertical axis. Recommended tags are those placed above 0.29.

	Before filtering		After filtering	
	FREESOUND	FLICKR1M	FREESOUND	FLICKR1M
Number of resources	118,629	107,617	118,629	107,617
Number of unique tags[a]	33,790	27,969	6,232	5,760
Number of contributor users[b]	5,523	5,463	5,523	5,463
Number of tag applications	782,526	927,473	730,417	882,616

[a] Not necessarily semantically unique.
[b] Users that have contributed by uploading, at least, one resource.

Table 3.2: Basic statistics of the folksonomies of FREESOUND and FLICKR1M datasets. We see that the datasets feature comparable numbers. The numbers under the "After filtering" column are computed by only considering tags that appear in at least 10 different resources (see below).

point where it crosses the vertical axis (Fig. 3.5). The higher 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 this strategy is also parameter-free.

3.3 Evaluation

From the combination of the different strategies above, we can define several tag recommendation methods which we evaluate through a tag prediction task (Sec. 2.3.4). 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 that we use for that evaluation.

3.3.1 Datasets

We use two real-world datasets collected from the tagging systems of Freesound and Flickr. In the case of Freesound, we consider all user annotations between April 2005 and September 2011, directly extracted from the Freesound database. 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 January 2004 and December 2009. Flickr data was collected by Papadopoulos et al. (2010) and provided to us by the authors. To avoid confusion with the totality of the Flickr content, we will refer to the analysed Flickr subset as FLICKR1M. Table 3.2 shows some basic statistics about the folksonomies of both datasets.

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[27]. As opposed to other well-

[27]Since a software upgrade in 2011, Freesound requires a minimum of three tags to annot-

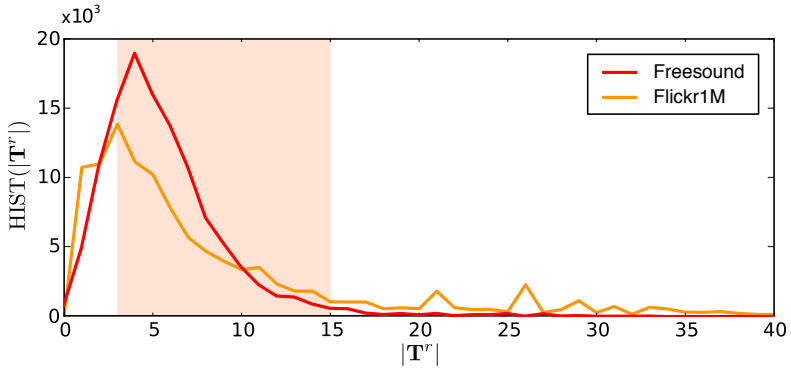


Figure 3.6: Histogram of the number of tags per resource $|\mathbf{T}^r|$ in FREESOUND and FLICKR1M. The average number of tags (standard deviation in parenthesis) per resource is 6.53 (6.47) and 7.50 (8.61) for FREESOUND and FLICKR1M, respectively.

studied 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 (Sec. 1.2). Hence, we can not weigh the association between a particular tag and a resource by the number of times the same association has been performed by different users.

The histogram of the number of tags per resource is qualitatively similar for the two datasets (Fig. 3.6). We are particularly interested in recommending tags for resources that fall in the range of $|\mathbf{T}^r| = [3, 15]$ tags, which are more than 80% and 65% of the total resources in FREESOUND and FLICKR1M, respectively (Fig. 3.6; 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 three tags as enough input information for our method to provide good recommendations. For resources with less than three tags, content-based strategies such as the ones outlined in Sec. 2.3.1 are probably more suited. On the other hand, we intuitively consider that resources with more than 15 tags are, in general, well enough described.

Among the set of all unique tags present in FREESOUND and FLICKR1M folksonomies, we apply a threshold $\omega = 10$ 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 (Sec. 3.2.1). After applying this threshold, we are left with 6,232 unique

ate 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 large enough number not to be considered as a restriction for normal tagging behaviour.

tags in the FREESOUND folksonomy (representing approximately 20% of the total) and with 5,760 unique tags in FLICKR1M (also representing approximately 20% of the total). This also means that we filter out all tag applications that do not associate any of these selected tags. Importantly, approximately 90% of tag applications in both FREESOUND and FLICKR1M involve one of these tags, thus we still take into account the vast majority of the original information (Table 3.2).

3.3.2 Methodology

Our evaluation methodology follows a standard information retrieval prediction task 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. We mentioned this fact in Sec. 2.3.4 and further discuss it in Sec. 3.5.

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 $\mathcal{S}$ 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 applications involving resources from the training set). For each resource in the evaluation set, we randomly delete a set of tags $\mathbf{T}_{\mathrm{D}}$ from its originally assigned tags, yielding $\mathbf{T}_{\mathrm{I}}$, the input to our system. The number of tags we delete is chosen uniformly at random, with the only constraint that the length of $\mathbf{T}_{\mathrm{I}}$ must be maintained in the range of $|\mathbf{T}_{\mathrm{I}}| = [3, 15]$ (see previous section). This constraint also implies that, in order to be able to remove at least one tag for each resource ($|\mathbf{T}_{\mathrm{D}}| \geq 1$), we can only consider for evaluation resources with 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.

Regarding evaluation measures, we compute P, R and F as defined in Eq. 2.1 (Sec. 2.3.4). 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. In addition to P, R and F, for each individual resource we also measure the number of recommended tags $|\mathbf{T}_{\mathrm{R}}|$. Evaluating $|\mathbf{T}_{\mathrm{R}}|$ is important because the longer the recommendation, the more comprehensive it potentially is, and the more difficult it is to maintain high precision values. We further discuss this aspect in Sec. 3.5.

Name	Aggregation step	Selection step
	Tag recommendation methods	
SimP@α	Similarity-based	Percentage ($\alpha = 0.30$[a], $\alpha = 0.20$[b])
SimST	Similarity-based	Statistical Test
SimKP@β	Similarity-based	Kernel Percentage ($\beta = 0.005$)
SimLR	Similarity-based	Linear Regression
RankP@α	Rank-based	Percentage ($\alpha = 0.15$[a], $\alpha = 0.10$[b])
RankST	Rank-based	Statistical Test
RankKP@β	Rank-based	Kernel Percentage ($\beta = 0.01$)
RankLR	Rank-based	Linear Regression
	Baseline methods	
BRankFIX@κ	Rank-based	Fixed number ($\kappa \in [1, 10]$)
BSimFIX@κ	Similarity-based	Fixed number ($\kappa \in [1, 10]$)
BRepeated@ϱ	Repeated tags in all sets $\mathbf{T}^i_{\mathrm{C}}(\varrho \in [2, 10])$	
BRandom	Random replacement of $\mathbf{T}_{\mathrm{R}}$.	
	State of the art baseline methods	
GW@κ	Garg & Weber (2008)	Fixed number ($\kappa \in [1, 10]$)
SZ@κ	Sigurbjörnsson & Zwol (2008)	Fixed number ($\kappa \in [1, 10]$)

[a] Parameter settings for FREESOUND estimated in preliminary experiments.
[b] Parameter settings for FLICKR1M estimated in preliminary experiments.

Table 3.3: Evaluated tag recommendation methods. All methods are evaluated using cosine similarity and $\theta = 100$.

A general characterisation of the number of recommended tags per method is also obtained by averaging $|\mathbf{T}_{\mathrm{R}}|$ across all resources evaluated with a particular recommendation method.

Table 3.3 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 propose. To avoid an intractable number of possible combinations, all methods are evaluated using only cosine similarity for Step 1, and setting $\theta = 100$ (getting a maximum of 100 candidates for each input tag). We choose cosine similarity as default because of its widespread usage in the literature, and $\theta = 100$ as an intuitively big enough number of candidates per input tag. We later study the influence of the chosen similarity measure and θ, 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 subset of 10,000 resources from the main evaluation 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, Table 3.3) are simpler versions of the proposed methods that we use for comparative purposes. On the one hand, we compare with two methods that implement a very simple strategy for selecting which tags to recommend (Step 3) and always recommend the first κ tags from $\mathbf{T}_{\mathrm{A}}$, sorted by their scores (BRankFIX@κ and BSimFIX@κ). We run these algorithms for values of κ 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 when fixing the number of tags to recommend. In preliminary experiments, we qualitatively observed a clear decrease of performance for values of κ close to 10, therefore values $\kappa > 10$ are not considered (this also applies to other methods that have κ as a parameter, see below). On the other hand, we compare with an even simpler method (BRepeated@ϱ) which, considering the union of all sets of candidates $\mathbf{T}_{\mathrm{C}}^1, \mathbf{T}_{\mathrm{C}}^2, \ldots\ \mathbf{T}_{\mathrm{C}}^i$ for a given resource, only recommends tags that are repeated more than ϱ times (independently of their scores). We run this algorithm for values of ϱ 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 $\mathbf{T}_{\mathrm{R}}$ with a random selection (of the same length) taken from $\mathbf{T}_{\mathrm{A}}$. 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 $\mathbf{T}_{\mathrm{R}}$. 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 higher F. Notice however, that these recommendations are not totally random: recommended tags are chosen from $\mathbf{T}_{\mathrm{A}}$, 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, Table 3.3) correspond to our implementations of the tag recommendation methods described by 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 κ recommended tags ranging from 1 to 10 (and only report the best result found). Garg & Weber (2008) describe several methods which contain different degrees of user personalisation. We implemented the "global" method which is not personalised and thus can be meaningfully compared to our methods. We implemented GW and SZ following the original references and set their parameters accordingly.

3.4 Results

3.4.1 Recommendation accuracy

From the average P, R and F values for each one of the evaluated methods using the FREESOUND and FLICKR1M datasets, we observe that Rank-based methods generally report higher F than Similarity-based methods (Tables 3.4 and 3.5). 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 assessed the statistical significance of this increase by performing pairwise Kruskal-Wallis tests (Corder & Foreman, 2009) between the results of each Rank-based method and its Similarity-based counterpart, and all have shown to be statistically significant[28], with a p-value several orders of magnitude below 0.01 (denoted as $p \ll 0.01$). These results indicate that Step 2 (Aggregation of candidate tags) is better accomplished using the Rank-based Strategy.

Regarding the results of the different strategies for Step 3 (Selection of tags to recommend), we observe a very similar behaviour in FREESOUND and FLICKR-1M (Tables 3.4 and 3.5, respectively). 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 for FREESOUND ($p \ll 0.01$), and 0.048 for FLICKR1M ($p \ll 0.01$). 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 aggregation provides the best obtained results in both datasets. When compared to the other selection strategies with Rank-based aggregation, it reports an average F increase of 0.025 for FREESOUND ($p \ll 0.01$), and 0.039 for FLICKR1M ($p \ll 0.01$). The similar results observed with FREESOUND and FLICKR1M partially support the idea that the proposed methods are generalisable to different kinds of data.

Having a look at the results of the baseline methods based on recommending a fixed number of 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, and that their reported results correspond to the upper bound of their performance (Sec. 3.3.2). That good performance thus 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. If we compare these baseline methods

[28] In the rest of this chapter, 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.

	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*

Table 3.4: Average precision P, recall R and f-measure F for tag recommendation methods using the Freesound dataset, sorted by f-measure. Baseline methods are marked in italics. For the sake of readability, we only show the results of baseline methods for the values of κ and ϱ that reported higher f-measure.

	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*

Table 3.5: Average precision P, recall R and f-measure F for tag recommendation methods using the Flickr1M dataset, sorted by f-measure. Baseline methods are marked in italics. For the sake of readability, we only show the results of baseline methods for the values of κ and ϱ that reported higher f-measure.

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@ϱ reports very low results both in FREESOUND and FLICKR1M datasets, and BRandom baseline remains significantly below all the other methods.

3.4.2 Number of recommended tags

Another aspect to evaluate from the tag recommendation methods is the number of tags that they recommend $|\mathbf{T}_\mathrm{R}|$. Table 3.6 shows the average $|\mathbf{T}_\mathrm{R}|$ 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 best scoring methods, corresponding to the first positions of the table, recommend more tags than BRankFIX@2 and GW@2 (Table 3.6), and at the same time report higher (or very similar) precision values and overall f-measure (see Tables 3.4 and 3.5). If we look at the evaluation results obtained with BRankFIX@κ methods when recommending more than two tags, we observe significant drops in precision ($P = 0.323$ for $\kappa = 3$ and $P = 0.272$ for $\kappa = 4$ in FREESOUND, and $P = 0.391$ for $\kappa = 3$ and $P = 0.333$ for $\kappa = 4$ in FLICKR1M). Similar precision drops are observed in GW@κ ($P = 0.306$ for $\kappa = 3$ and $P = 0.257$ for $\kappa = 4$ in FREESOUND, and $P = 0.396$ for $\kappa = 3$ and $P = 0.340$ for $\kappa = 4$ in FLICKR1M). This further highlights the superiority of our proposed methods over the baselines.

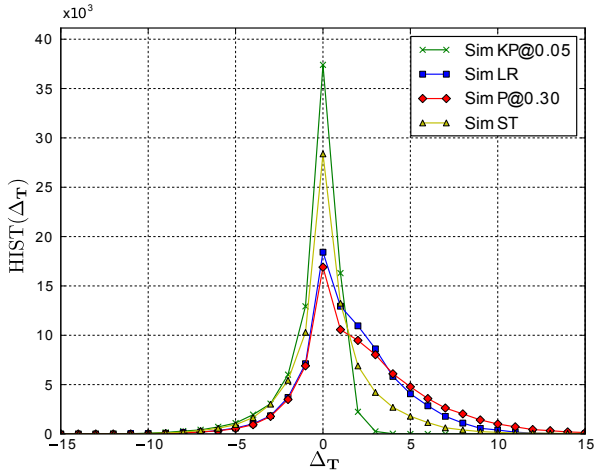
It is also interesting to see that the number of recommended tags is not only driven by the selection strategy of Step 3, but also depends on the type of aggregation used in Step 2. Both in FREESOUND and FLICKR1M, we observe that when using Rank-based aggregation, highest $|\mathbf{T}_\mathrm{R}|$ is obtained using the strategy of Linear Regression for selecting which tags to recommend (followed by Statistical Test, Percentage and Kernel Percentage strategies). However, when using Similarity-based aggregation, the highest $|\mathbf{T}_\mathrm{R}|$ is obtained with the Percentage Strategy, followed by Linear Regression, Statistical Test and Kernel Percentage strategies (Table 3.6). This shows that the selection strategies behave differently if the scores of $\mathbf{T}_\mathrm{A}$ are ranks or similarity values. In general, Rank-based methods recommend more tags than their Similarity-based counterparts, with an average $|\mathbf{T}_\mathrm{R}|$ increase of 0.38 for FREESOUND ($p \ll 0.01$), and 0.86 for FLICKR1M ($p \ll 0.01$). Given that Rank-based aggregation methods also report higher F, this reinforces the aforementioned observation that Step 2 is better accomplished using the Rank-based Strategy.

Furthermore, we also looked at the difference between the number of recommended tags and the number of tags that are deleted for each resource

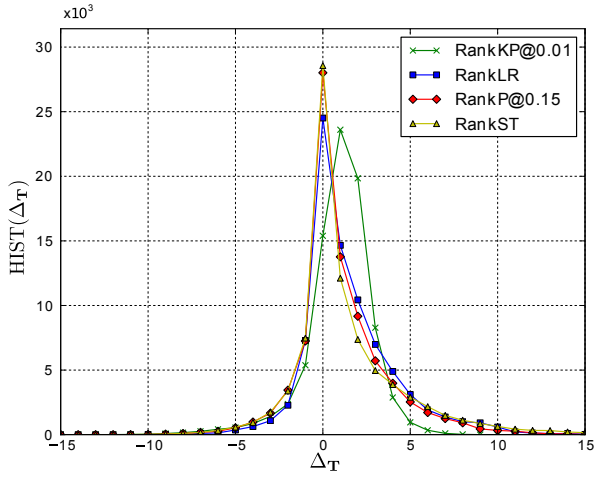
FREESOUND		FLICKR1M	
Method	$\|\mathbf{T_R}\|$	**Method**	$\|\mathbf{T_R}\|$
RankP@0.15	3.03 (2.60)	RankP@0.10	2.68 (1.96)
RankST	3.36 (3.30)	*GW@2*	*2.00 (0.00)*
RankLR	3.55 (7.14)	*BRankFIX@2*	*2.00 (0.00)*
BRankFIX@2	*2.00 (0.00)*	RankST	3.96 (3.64)
RankKP@0.01	2.89 (1.29)	RankLR	4.64 (4.25)
GW@2	*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)*

Table 3.6: Average number of recommended tags $|\mathbf{T}_\mathrm{R}|$ for tag recommendation methods using the FREESOUND and FLICKR1M datasets (standard deviation into parentheses). Methods are displayed and sorted according to the F values of Tables 3.4 and 3.5. Baseline methods are marked in italics.

($\Delta_\mathbf{T} = |\mathbf{T}_\mathrm{R}| - |\mathbf{T}_\mathrm{D}|$). In Fig. 3.7 we show the histogram of $\Delta_\mathbf{T}$ for our proposed methods. We observe that most of our proposed methods report the maximum peak of the histogram at $\Delta_\mathbf{T} = 0$ (Fig. 3.7). 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. 3.8). We can qualitatively observe how $|\mathbf{T}_\mathrm{R}|$ grows along with $|\mathbf{T}_\mathrm{D}|$, specially for low $|\mathbf{T}_\mathrm{I}|$. It can also be observed that there is a tendency of $|\mathbf{T}_\mathrm{R}|$ increasing when $|\mathbf{T}_\mathrm{I}|$ 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).



(a)



(b)

Figure 3.7: Histogram of the difference between the number of recommended tags and the number of deleted tags $\Delta_{\mathbf{T}}$ for Similarity-based (a) and Rank-based (b) tag recommendation methods using FREESOUND dataset. Qualitatively similar results were obtained with FLICKR1M.

3.4.3 Other relevant aspects

In order to better understand the behaviour of the proposed tag recommendation methods, we have carried out further analyses on the influence 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 3.3). In the following sections we report experiments concerning these aspects.

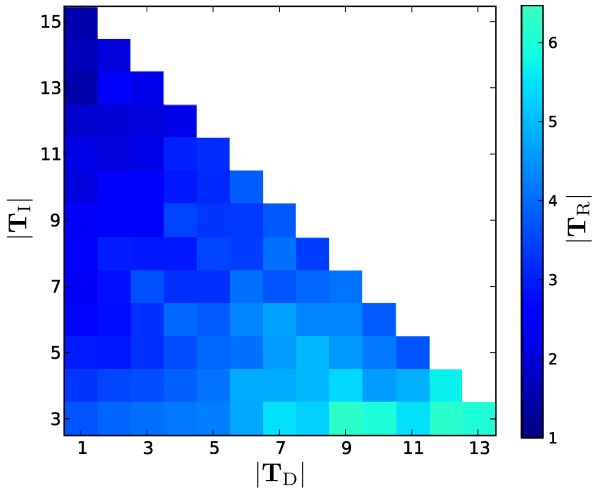


Figure 3.8: Average number of recommended tags $|\mathbf{T}_\mathrm{R}|$ as a function of the number of input tags $|\mathbf{T}_\mathrm{I}|$ and the number of deleted tags $|\mathbf{T}_\mathrm{D}|$, for method RankST and FREESOUND dataset.

Limiting the minimum number of input tags

To assess the influence of limiting the number of input tags, we now repeat the main experiments but include resources evaluated with less than three input tags. As it could be expected, we obtain lower F scores (Table 3.7). On 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. 3.9 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 3.4 and 3.5 emphasizes that, even outside the optimum conditions, the proposed methods are still useful to some extent.

Using alternative similarity measures

As has been explained in the evaluation methodology, all previously reported experiments have been performed using cosine similarity as the similarity measure for Step 1. In this subsection we repeat the evaluation for the best scoring methods but now using Jaccard and tag co-occurrence as similarity measures (Table 3.8). In both datasets and for all methods, cosine similarity is

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

Table 3.7: 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.

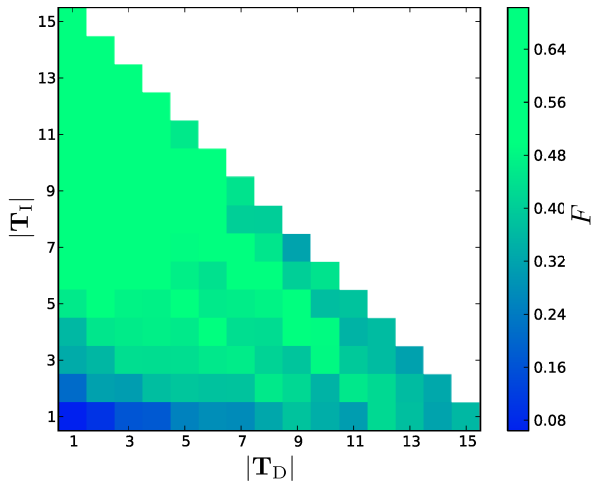


Figure 3.9: Average f-measure F as a function of the number of input tags $|\mathbf{T}_\mathrm{I}|$ and the number of deleted tags $|\mathbf{T}_\mathrm{D}|$ for method RankP@0.15 and FREESOUND dataset. This plot includes the results of resources evaluated with fewer than three input tags.

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 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, 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

In order to understand the effect of the number of candidates per input tag θ (Step 1), we have performed a series of experiments with the best scoring methods. Similar to the main experiments described in Sec. 3.3.2, we have performed 10-fold cross validations for each one of the best scoring methods, giving different values to θ. 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 $\theta = [75, 150]$, and then show a very slow decaying tendency (Fig. 3.10). 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 $\theta = [125, 400]$). These results suggest that after selecting a sufficient amount of θ candidates for each input tag, the most relevant tags have already been selected, and increasing θ does not have a relevant impact on the output of the recommendation as score values for the "extra" candidates are generally low. According to Fig. 3.10, for most of the methods, highest F is obtained with $\theta \approx 125$, which is slightly higher than the value we used for our main experiments ($\theta = 100$). However, the average F increase is less than 1%, and significance tests fail with $p \approx 0.10$ when comparing the methods configurations with $\theta = 100$ and $\theta = 125$.

Contribution of each step of the recommendation scheme

To finish our analysis, we perform several experiments to evaluate the contribution of each step of the proposed tag recommendation scheme. For the best scoring methods, we have repeated the 10-fold cross validations of the

Method	Precision	Recall	F-measure	$\|\mathbf{T_R}\|$
		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 3.8: Average precision P, recall R, f-measure F and number of recommended tags $|\mathbf{T}_\mathrm{R}|$, using different similarity measures.

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 θ random candidates from the whole vocabulary of the folksonomy (using $\theta = 100$). 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 $\mathbf{T}_\mathrm{A}$, 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 κ tags from $\mathbf{T}_\mathrm{A}$ (sorted by the scores of candidates). In that case, κ is determined by a random number generator with a normal distribution with the same mean (μ) and standard deviation (σ) as that observed for the number of

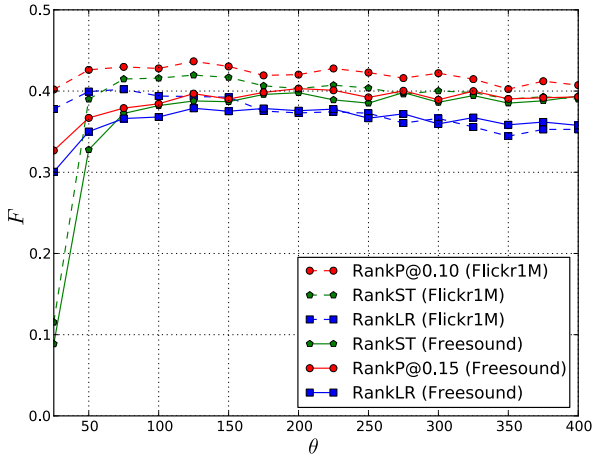


Figure 3.10: Average f-measure F with different values of θ for the best scoring recommendation methods in FREESOUND and FLICKR1M (each experiment performed with 10,000 resources).

deleted tags in the main experiments ($\mu = 1.92$ 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 $\Delta_{\mathbf{T}} \approx 0$.

Runs 1 and 2 report very low F in both datasets (Table 3.9). 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 3.9). Given that Steps 1 and 2 are tightly coupled, failing in any of them has a very important 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 with 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.

Method	Run 1	Run 2	Run 3	No rand.
		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

Table 3.9: Average f-measures F after randomising steps 1, 2 and 3 of the best scoring tag recommendation methods in FREESOUND and FLICKR1M. The "No rand." column shows the performance of the recommendation methods when no steps are randomised.

3.5 Conclusion and discussion

In this chapter we have presented a general scheme for tag recommendation systems based on tag co-occurrence in folksonomies. This scheme is composed of three steps for which we have proposed different strategies. Step 1, Candidate tag selection, selects a number of candidate tags for every input tag based on a tag-tag similarity matrix derived from a folksonomy. Three variants of this step are given by the usage of alternative similarity measures. Step 2, Aggregation of candidate tags, assigns scores to the candidates from Step 1 and merges them all in a single list of candidate tags. For this step, we have proposed two strategies which differ in the way scores are assigned. Finally, Step 3, Selection of 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 different online sharing platforms. The main bottleneck in terms of scalability lies in the computation of the tag-tag similarity matrix that informs the candidate selection step. However, this matrix can be computed offline, and its size can be easily reduced by raising the threshold ω during the construction of the association matrix. This means that the described recommendation methods can scale well to even bigger amounts of data, as the number of unique tags above the threshold ω grows much more slowly than the number of resources. Hence, the simplicity of the described methods makes them suitable for dealing with large-scale datasets such as the ones we have used here. Moreover, the described tag recommendation methods are easily

adaptable to any other 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 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 qualitatively 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 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 (Garg & Weber, 2008). To exemplify this observation, Table 3.10 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. 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.

Overall, the work we have carried out shows that the proposed scheme for folksonomy-based tag recommendation can successfully be used for recommending tags to online resources. In the following chapter, we propose an improvement for the best-scoring recommendation method described here, and carry out an evaluation based on user assessment of the recommended tags.

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

Table 3.10: 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].

CHAPTER 4

An enhancement: class-based tag recommendation

4.1 Introduction

In this chapter we build on the tag recommendation methods described in the previous chapter and extend them. Furthermore, we perform an evaluation of the extended recommendation system through an online experiment with real users. Hence, the contribution of the present chapter is twofold. Firstly, we propose an extended version of the best performing tag recommendation method described in Chapter 3 (RankP@α). 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, and the extended tag recommendation system incorporates some domain-specific knowledge in the form of resource categories. We evaluate the automatic classification process separately from the rest of the tag recommendation system.

Secondly, we perform a comprehensive user-based evaluation through an online experiment. In it, participants are presented with some resources which have to be annotated with the help of the tag recommendation system. These kinds of user-based evaluations are very costly, and we have seen that they are not very common in the tag recommendation literature (Sec. 2.3.4). 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 such as the one described in the previous chapter (Sec. 3.3, also see Sec. 2.3.4). In the previous chapter, we evaluated the tag recommendation methods in the audio and image domains (i.e., using data from Freesound and Flickr),

and obtained similar results in both scenarios. In this chapter, due to the extended recommendation method being based on the automatic classification of resources, evaluations are solely carried out in the context of Freesound.

The rest of this chapter is organised as follows. First, in Sec. 4.2, we describe the extended tag recommendation method based on the classification of input tags. Then, we evaluate the classifier that we use in the recommendation process of the extended method (Sec. 4.3). In Sec. 4.4, we describe our main evaluation methodology based on the online experiment, and report its results in Sec. 4.5. The complementary prediction-based evaluation of the recommendation methods is performed in Sec. 4.6. Finally, we conclude the chapter with a discussion about our findings and future work (Sec. 4.7).

4.2 Methods

Similar to the tag recommendation methods described in the previous chapter, the tag recommendation method described here is entirely based on tag-tag similarities derived from the folksonomy of Freesound. We begin by summarising the steps of the best performing tag recommendation method described in the previous chapter (RankP@α, with $\alpha = 0.15$ for FREESOUND dataset). We refer to this method as the "general" method or GEN for short. Then, we describe the extended parts of the new method, which mainly include a class detection step, and the computation of tag-tag similarity matrices based on that classification. We refer to the extended method as "class-based" or CLA for short.

4.2.1 General tag recommendation

Given a set of input tags $\mathbf{T}_\mathrm{I}$ and a tag-tag similarity matrix $\mathcal{S}$, the GEN method can generate a sorted list of recommended tags $\mathbf{T}_\mathrm{R}$. It consists of the three steps depicted in the top of Fig. 4.1:

1. Candidate tag selection: Given a set of input tags $\mathbf{T}_\mathrm{I}$, this step uses a tag-tag similarity matrix $\mathcal{S}$ derived from the Freesound folksonomy to select a set of θ candidate tags $\mathbf{T}_\mathrm{C}^i$ for each input tag $\mathbf{T}_{\mathrm{I}_i}$. 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 sounds in the Freesound folksonomy ($d_{i,j} = 1$ if sound r_i is labeled with tag t_j, and $d_{i,j} = 0$ otherwise). Given $\mathcal{D}$, the tag-tag similarity matrix is obtained as $\mathcal{S} = \mathcal{D}\mathcal{D}'$, 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 sounds. Tags in $\mathbf{T}_\mathrm{C}^i$ are selected as the $\theta = 100$ most similar tags to a given input tag $\mathbf{T}_{\mathrm{I}_i}$.

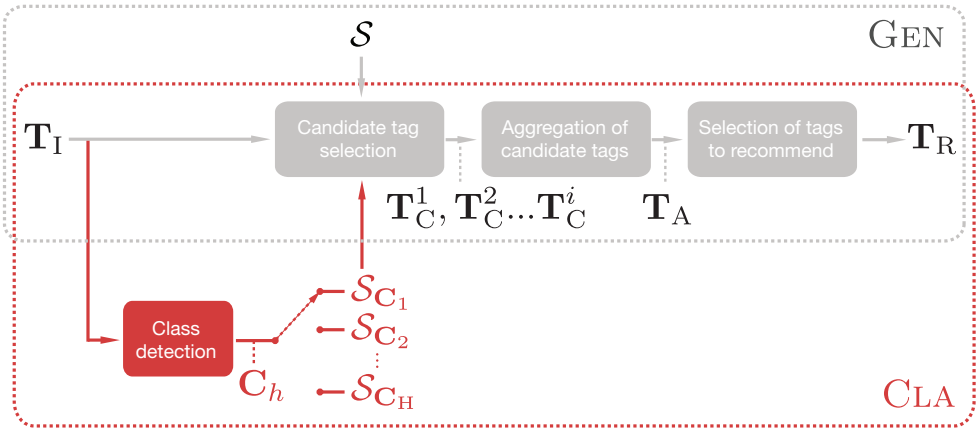


Figure 4.1: Block diagram of the general (GEN) and class-based (CLA) tag recommendation methods.

2. Aggregation of candidate tags: Given the sets $\mathbf{T}_\mathrm{C}^i$ from the first step, candidate tags are assigned a score ϕ and aggregated into a single list of tags with scores $\mathbf{T}_\mathrm{A}$. Such a score is determined by the candidate similarity-based ranking so that $\phi = 1$ for the most dissimilar candidate to a given input tag and $\phi = \theta$ for the most similar one. The scores of tags that are present in different sets of candidates $\mathbf{T}_\mathrm{C}^i$ are added when aggregated in the final set $\mathbf{T}_\mathrm{A}$.

3. Selection of tags to recommend: Considering the scores in $\mathbf{T}_\mathrm{A}$, this step 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 $\mathbf{T}_\mathrm{A}$, and use a percentage parameter of $\alpha = 0.15$ (Sec. 3.2.3). Tags in $\mathbf{T}_\mathrm{A}$ are sorted by their score and those that satisfy $\phi \geq \varepsilon$ are outputted as $\mathbf{T}_\mathrm{R}$, the final set of recommended tags.

4.2.2 Class-based tag recommendation

The proposed class-based tag recommendation method is a variation of GEN based on the classification of the input tags $\mathbf{T}_\mathrm{I}$ into a set of H predefined audio classes[29]. For every class $\mathbf{C}_h$ ($\mathbf{C}_h \in \mathbf{C}$, where $\mathbf{C}$ is the set of defined audio classes), a tag-tag similarity matrix $\mathcal{S}_{\mathbf{C}_h}$ is built in the same way as in the GEN method, except that in this case only the information of tag applications involving the sounds 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 $\mathcal{S}_{\mathbf{C}_h}$ that is used in the candidate tag selection step of the recommendation process depends on the classification of the input tags $\mathbf{T}_\mathrm{I}$ (Fig. 4.1). Once the candidates are selected, the other two steps (Aggregation

[29] When referring to audio classes, we may use the terms "class" or "category" indistinctly.

of candidate tags and Selection of tags to recommend) are computed in exactly the same way as in GEN. We now describe the classification system that we use in the class detection step, and then explain the computation of the tag-tag similarity matrices.

Classification system

In order to classify a set of input tags $\mathbf{T}_\mathrm{I}$ of a sound r into one of the audio classes, we make use of standard machine learning techniques. The structure of the classification system is very similar to what can be found in the existing literature on the classification of sound effects (Kuo & Zhang, 1999; Casey, 2002; Sundaram & Narayanan, 2008; Roma et al., 2010), musical instruments (Herrera et al., 2003; Livshin et al., 2003; Cano et al., 2005), or music genre and mood (Laurier et al., 2008; Bischoff et al., 2009; Chen et al., 2009; Tao et al., 2010). In these works, a set of low-level audio features is typically extracted from sounds in a given collection, yielding a feature vector representation of every sound. Also, sounds are manually annotated using the concepts of a taxonomy representing the particular classification domain (e.g., a taxonomy of musical instruments or sound effects). These taxonomies tend to be rather small, typically including less than 20 concepts. Then, supervised learning is performed using a classifier trained with the feature vectors corresponding to annotated sounds.

The classification we perform here only differs from this scheme in that we do not extract audio features from the sounds we classify, but use instead their existing associated tags (i.e., taglines) as feature vectors to train the classifier. To do this, we follow a bag-of-words approach where each sound is represented as a vector whose elements indicate the presence or absence of a particular tag. Feature vectors contain all possible tags in the collection, thus their dimensionality is very high. Here we do not carry on any dimensionality reduction step to lower the size of the feature vectors. Instead, in order to keep them in manageable sizes, we apply the same threshold $\omega = 10$ described in Secs. 3.3.1 and 4.2.1, removing all tags that are used less than 10 times. The resulting feature vectors are very sparse, which makes the problem similar to what is normally found in text classification, where high dimensionality and sparseness are commonplace (Sebastiani, 2002). We experiment with a support vector machine (SVM) and a naive Bayes (NB) classifier, as these have been shown to be well suited for high dimensional and sparse classification tasks such as the one we are facing here (Bennett & Campbell, 2000; Sebastiani, 2002)[30]. By not including extracted audio features in the classification system, we maintain a certain degree of generalisability for the class-based tag

[30] We implement the classifiers using the "scikit-learn" Python package (`http://scikit-learn.org`). We use the classes `LinearSVC` and `BernoulliNB` for SVM and NB, respectively, with default parameters. `LinearSVC` follows the "one versus all" approach for multiclass classification.

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 sounds, 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.
Voice	Various voice-related sounds such as text reading, single words or recordings of text-to-speech processors.

Table 4.1: Name and descriptions of the audio classes we defined.

recommendation method, as the methodology remains applicable to other domains without further modifications. Nevertheless, what necessarily changes from domain to domain is the definition of classes.

Given the heterogeneity of the audio content in Freesound, we define the audio categories that we want to detect in a way that these can virtually include the whole range of sounds that can be found in Freesound. Hence, we define a total of $H = 5$ audio categories in which sounds can be classified. The resulting categories, shown in Table 4.1, are quite general and are in line with other sound categorisations reported in the literature (Casey, 2002; Roma et al., 2010). In order to create a dataset for the supervised learning process, we manually assigned one of the above categories to a number of sounds from Freesound. To do this, we followed an iterative process in which we were presented with randomly chosen sounds from Freesound, and assigned them to one of the five categories. As it can be imagined, these categories are not completely orthogonal, and there were sounds for which the decision was not straightforward just by listening to the audio. In these cases, we also relied on provided textual descriptions of the sounds (i.e., their textual descriptions in Freesound). The crafted dataset includes a minimum of 2,088 sounds per category (corresponding to the case of Speech) and a maximum of 6,341 (for the case of Samples). Comparing the totality of Freesound sounds and the manually annotated subset, we observe qualitatively similar relative distributions of tag occurrences and number of tags per sound. Fig. 4.2 shows the most commonly used tags for the five defined audio categories. Using the manually crafted ground truth and the vector representations of sounds according to their taglines, we can train a classifier which, given a set of input tags $\mathbf{T}_\mathrm{I}$, can predict which category

$\mathbf{C}_h$ better fits the input. In Sec. 4.3 we evaluate the classifier and report the results in terms of classification accuracy.

Computation of tag-tag similarity matrices

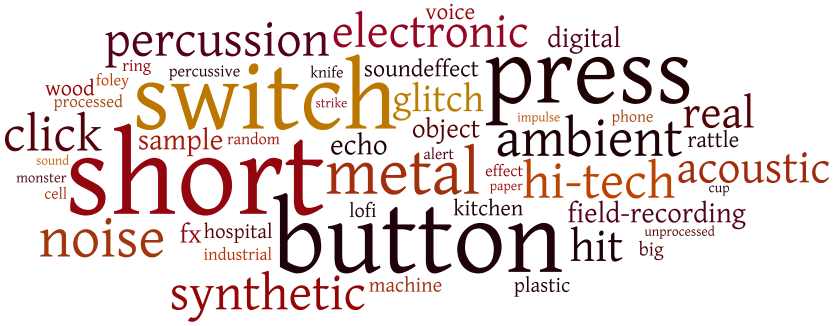
As mentioned, the process of building the tag-tag similarity matrices $\mathcal{S}_{\mathbf{C}_h}$ is the same as the one for building $\mathcal{S}$, except that for every matrix $\mathcal{S}_{\mathbf{C}_h}$ we only consider information about tag applications from sounds belonging to $\mathbf{C}_h$. For doing that, we reuse the classification system described above and classify all sounds of Freesound into one of the five audio classes, using as input tags the original taglines of the sounds in Freesound. Then, matrices $\mathcal{S}_{\mathbf{C}_h}$ can be built by only considering the columns of $\mathcal{D}$ corresponding to the sounds of $\mathbf{C}_h$. Hence, $\mathcal{S}_{\mathbf{C}_h} = \mathcal{D}_{\mathbf{C}_h}\mathcal{D}'_{\mathbf{C}_h}$, where $\mathcal{D}_{\mathbf{C}_h}$ is a subset of $\mathcal{D}$ in which the columns corresponding to sounds not in $\mathbf{C}_h$ are removed. Each matrix $\mathcal{S}_{\mathbf{C}_h}$ is normalised using the same process we use for $\mathcal{S}$ to obtain cosine similarity (Secs. 3.2.1 and 4.2.1).

Notice that the similarity value between two tags t_i and t_j will be different in every matrix $\mathcal{S}_{\mathbf{C}_h}$ and in $\mathcal{S}$, with $\mathcal{S}_{\mathbf{C}_h}$ being tailored to the particular context of the h-th class. Note also that the number of distinct tags resulting from considering all sounds belonging to $\mathbf{C}_h$ will be smaller than the total number of distinct tags resulting from considering all sounds 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 $\mathcal{S}_{\mathbf{C}_h}$, corresponding to the tags that are not used in the context of the particular class $\mathbf{C}_h$. These tags are thus never recommended when using $\mathcal{S}_{\mathbf{C}_h}$.

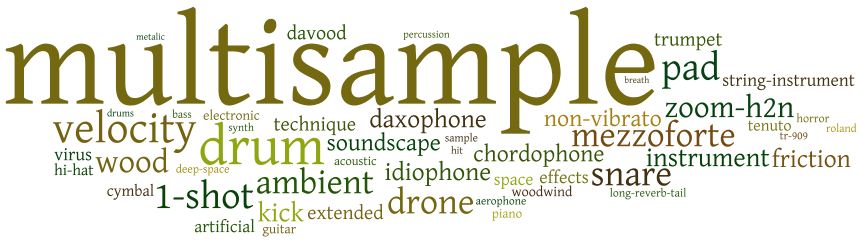
4.3 Results and evaluation of the classification system

4.3.1 Methodology

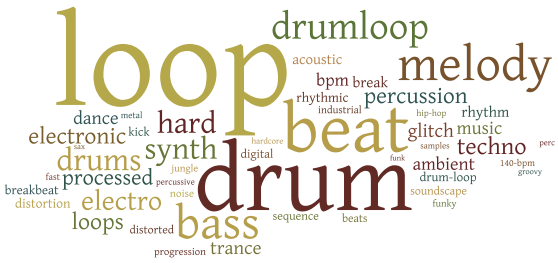
To evaluate the classification system we follow a random sub-sampling cross-validation strategy where we split the aforementioned ground truth (i.e., the dataset) into training and testing sets. We then compute the out-of-sample accuracy as the percentage of well-classified instances from the testing set when using the fit from the training set. This process is repeated 100 times for each classifier and parameter configuration that we test (see below), and overall accuracy is obtained by averaging over the results of all repetitions. In each repetition, our dataset is composed of a random selection of 1,000 sounds from every category, adding up to a total of 5,000. This way we maintain a balance in the number of sounds per category. In addition, to avoid potential bias of our classifier to the tagging conventions of a particular user, we impose the limit of not getting more than 50 sounds of the same category uploaded



a) SoundFX



b) Samples



c) Music

d) Voice

e) Soundscape

Figure 4.2: Tagclouds of the 50 most used tags in the five defined audio categories. The size of the tags is proportional to the frequency of occurrence among all sounds annotated under each category. For building these tagclouds, we only considered the set of sounds manually annotated as ground truth. Tagclouds were generated with the online tool available at `www.wordle.net`.

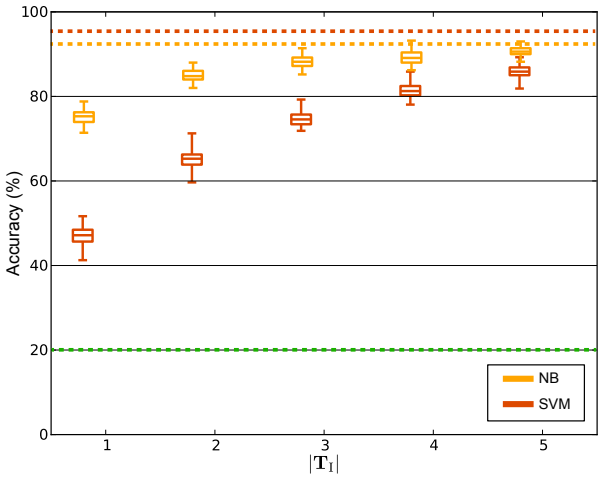


Figure 4.3: Classification accuracy using SVM and NB classifiers. The dashed line at 20% accuracy corresponds to the random baseline. The dashed lines around 95% (SVM) and 90% (NB) correspond to the accuracy achieved when no restriction on the number of tags for the testing set is performed, which can be considered as an upper bound limit.

by the same Freesound user. In each repetition, the testing set is selected as a random subset representing 10% of the data, and being equally-distributed among categories (i.e., 100 sounds per category).

As mentioned in Sec. 4.2.2, we test our method using SVM and NB classifiers. We also add a random classifier to serve as a baseline. Given that for the class-based tag recommendation method the classification step is extensively used in situations with few input tags (i.e., when users start introducing tags), we are specially interested in evaluating the accuracy of the classification system in conditions of tag scarcity (i.e., low $|\mathbf{T}_I|$). Hence, we introduce a limitation to the testing set consisting of randomly removing tags from sounds prior to classification, only leaving a particular number of N input tags per sound. We consider values of N ranging from 1 to 5. This obviously adds another constraint to the selection of the testing set, which is to make sure that selected sounds have at least N tags. The whole evaluation process is performed for all different values of N and for both SVM and NB classifiers, yielding a total of 10 evaluated experiment combinations.

4.3.2 Results

Fig. 4.3 shows the accuracies of our classification method for the experiment combinations described above. Note that all combinations are far above the random classifier accuracy of 20%. The NB classifier reports overall a higher accuracy than the SVM, with a statistically significant average accuracy increase of 10% ($p < 10^{-12}$). Statistical significance is assessed by considering

the maximum p-value across pairwise comparisons between experiment combinations and using the Wilcoxon signed-rank test with a significance level of 0.01 (Corder & Foreman, 2009). Overall, using the NB classifier, the classification system is able to successfully classify sounds among five generic categories inside the audio domain, with accuracies ranging from 75% to 90% depending on the number of input tags available for classification. As expected, the lowest accuracy is obtained when $|\mathbf{T}_\mathrm{I}| = 1$ (i.e., only one tag is given to the classifier). For $|\mathbf{T}_\mathrm{I}| \geq 4$, the classification accuracy reaches values close to 90%.

To complement these results, we performed additional experiments with different training set sizes (i.e., using less than than 90% of sounds for training). The results we obtained are consistent with those reported above with little variation on accuracy for training set percentages higher than 50%. This reinforces the validity of the classification results as the use of smaller training sets does not heavily affect classification accuracy. Furthermore, we also tested different values for the imposed maximum of 50 sounds uploaded by the same user in the same audio category. Our results show that the accuracy is not significantly influenced by such limit, thus asserting that the classifier is not biased to the tagging conventions of a particular user.

4.4 Evaluation of the tag recommendation methods

To evaluate the class-based tag recommendation method we designed an online experiment where participants had to tag a set of sounds from Freesound. The experiment was online for 15 days during June 2013, and was publicised in the Freesound front page. The goal of this experiment was twofold. First, we wanted to assess the usefulness of the Cla method with respect to the previous Gen method. Second, we wanted 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.

Along with Gen and Cla, in the experiment we also evaluated 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 $\mathbf{T}_\mathrm{R}$ 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 sounds. The dataset we used for the evaluation comprises Freesound data gathered between April 2005 and May 2012. Tables 4.2 and 4.3 show basic statistics of the dataset and

Freesound dataset	
Number of sounds	140,622
Number of unique tags[a]	43,696
Number of contributor users[b]	6,948
Number of tag applications	990,574
Average tags per sound (tagline length)	7.04

[a] Not necessarily semantically unique.
[b] Users that have contributed by uploading, at least, one resource.

Table 4.2: Basic statistics of the Freesound dataset (including data collected between April 2005 and May 2012).

Tag-tag similarity matrices		
	Num. sounds	**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 Voice	15,305	3,557

Table 4.3: Basic statistics of the resulting tag-tag similarity matrices.

the resulting tag-tag similarity matrices that we built with this dataset. The online-experiment proceeded as follows:

Instructions page: First, participants were presented with an introduction page displaying detailed instructions for the experiment (Fig. 4.4). Participants were told they would have to annotate 20 sounds from Freesound, using as many tags as they felt appropriate for every sound (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 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. 4.5) 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).

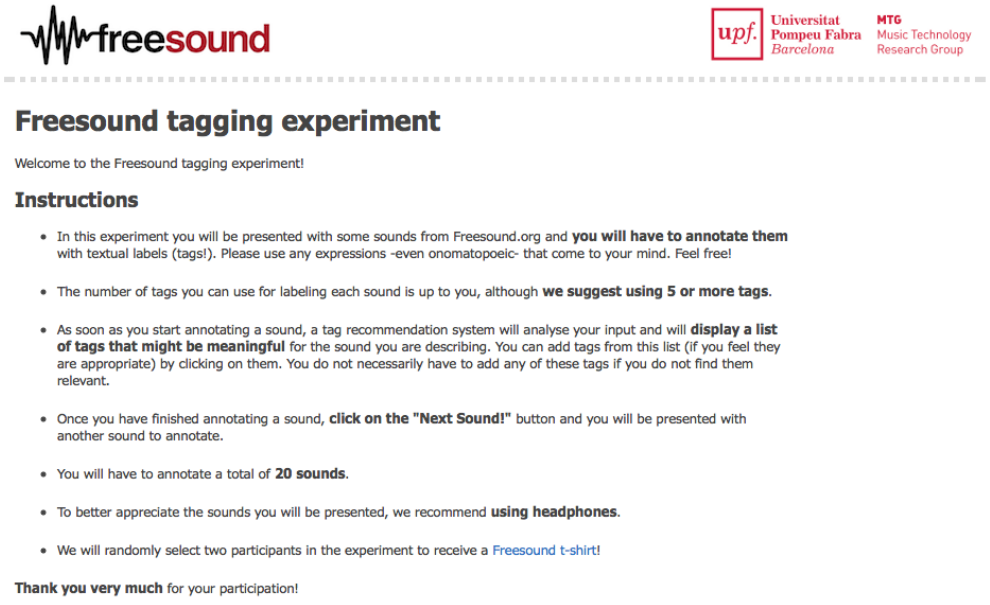


Figure 4.4: Screenshot of the instructions page.

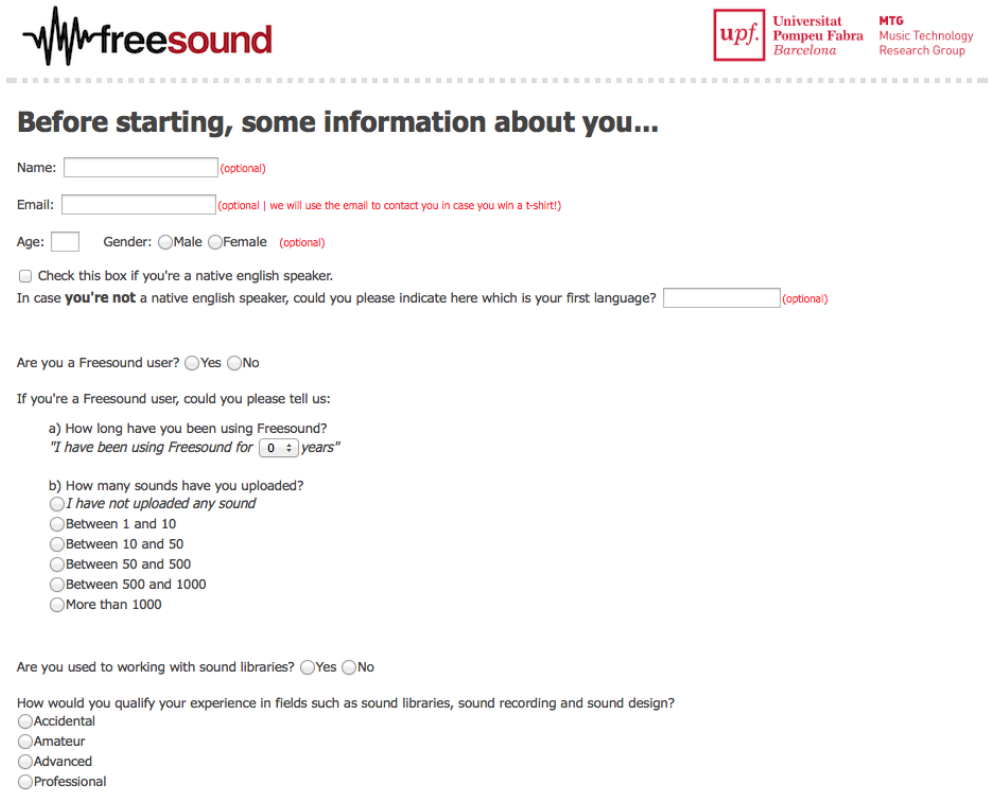


Figure 4.5: Screenshot of the questionnaire page.

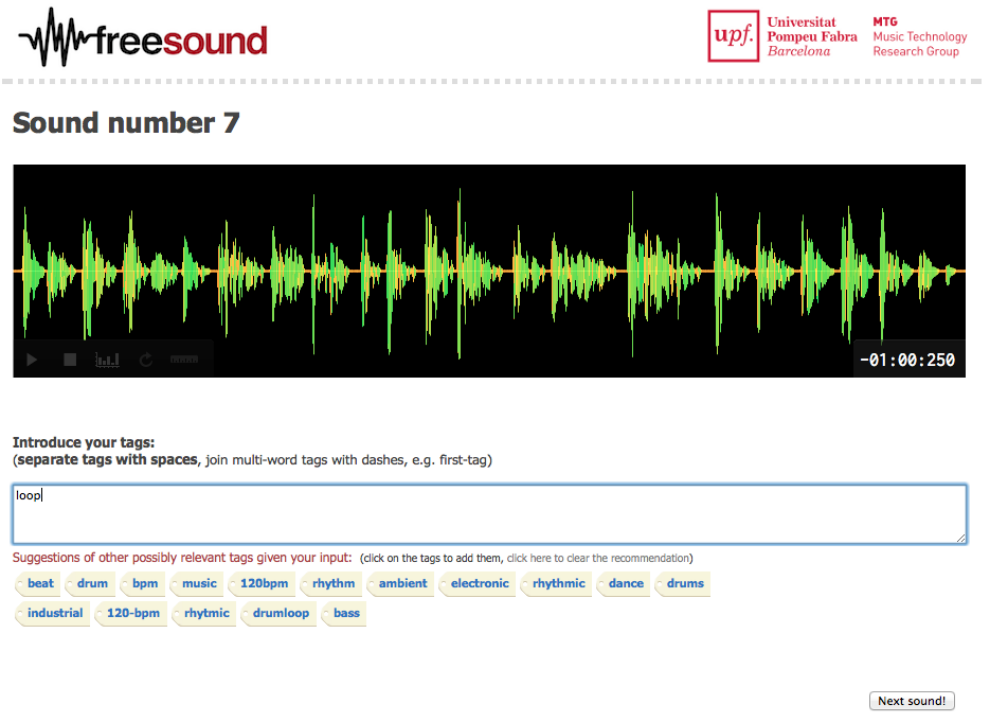


Figure 4.6: Screenshot of the sound annotation page.

Sound annotation: Once the questionnaire was completed, participants started annotating sounds. From the ground truth we defined when designing the recommendation system (Sec. 4.2.2), we manually selected 50 sounds per class[31]. These sounds 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 sounds, every participant was assigned a random selection of four sounds per class. Then, each of the four sounds was assigned a different tag recommendation method that would be used when the participant annotated the sound. In this way, every participant was assigned a total 20 sounds, equally distributed among audio classes and recommendation methods. Participants were presented with the first sound and had to annotate it by typing tags in a text box. The sound could be reproduced using a web player that also showed a visualisation of the waveform and the spectrogram of the sound (Fig. 4.6). 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 sound, and was being updated every time a new tag was written in the text box.

[31]The sounds 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 4.2.2.

Similar to the Freesound upload system, tags had to be separated by spaces and multi-words joined with hyphens. Hence, the recommendation was primarily updated every time a blank space was introduced. Users could click over the tags shown in the list to automatically append them in the text box (Fig. 4.6). Once a participant considered a sound was fully annotated, she could click on the "Next sound" button and be presented with the following sound. Participants were also provided an URL that they could save for later resuming the experiment in case they did not want to annotate all sounds 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 sounds, 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 thread of the Freesound forums.

Considering the logs gathered during the experiment, we define a simple measure for evaluating the "usefulness" of every tag recommendation method in the tagging process. The measure consists of counting, for every set of tags assigned to a sound 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 *correctly predicted* by the system or, what is the same, *accepted* by the participant). Then, this number can be averaged over all sounds annotated with each recommendation method, and obtain in this way a general characterisation of the method. Let $\mathbf{R}$ be a set of resources, let $\mathbf{T}^r$ be the set of tags that a participant used to annotate a particular sound r, and let $\mathbf{T}_{\mathrm{R}}^{r,m}$ 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 sound r. Then, we can define Ω, the number of correctly predicted tags (or accepted tags), as

$$\Omega = \frac{1}{|\mathbf{R}|} \sum_{r \in \mathbf{R}} \left| \mathbf{T}^r \cap \left(\bigcup_{m=1}^{M} \mathbf{T}_{\mathrm{R}}^{r,m} \right) \right|. \tag{4.1}$$

Notice that Ω is roughly equivalent to a standard recall measure (without the normalization term). We employ this measure instead of standard precision and recall (e.g., as done in Jäschke et al. 2009) because the nature of our evaluation has some particularities which make such metrics less useful. As described above, several tag recommendations are performed during the annotation of a single sound (i.e., every time that a new tag is introduced the recommendation is recomputed). As a result, the total number of recommended tags for every sound 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 sound with the final taglines assigned by users, we would obtain very low precision values which, in our opinion, are not as representative as Ω. 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 arguably more important than precision).

4.5 Results of the tag recommendation methods

During the two weeks the experiment was online, we gathered a total 201 experiment logs from 190 unique participants (a few 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 sounds 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 sound) and from experiments not reporting a minimum of three calls to the recommendation system for every annotated sound. We discard these logs as we consider that participants did not pay enough attention when annotating sounds 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 sections we report the results of the different aspects of the online experiment that we analyse.

4.5.1 Correctly predicted tags per recommendation method

First, we report the basic accuracy of the considered tag recommendation methods (Table 4.4, leftmost column). We observe that random methods RCla and RGen report considerably lower average Ω than Cla and Gen. Thus, our methods generate 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 found not to be statistically significant. In the following comparisons, and if not stated otherwise, statistical significance is assessed by performing pairwise comparisons using the Mann-Whitney U test with a significance level of 0.05 (Corder & Foreman, 2009).

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

	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 4.4: Average number of correctly predicted tags Ω (standard deviation in parenthesis) of the user-based evaluation approach for the different groups of participants.

beginning of the experiment (Table 4.4). In particular, we compute Ω 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 4.4), non-experienced participants (third column), native English speakers (fourth column), and non-native speakers (fifth column). We again observe that Cla reports higher averages than Gen, which 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 expert group (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 hand, 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, especially prominent 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 higher Ω, 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 correctly predicted tags among the native and non-native participant groups (Table 4.4). Native

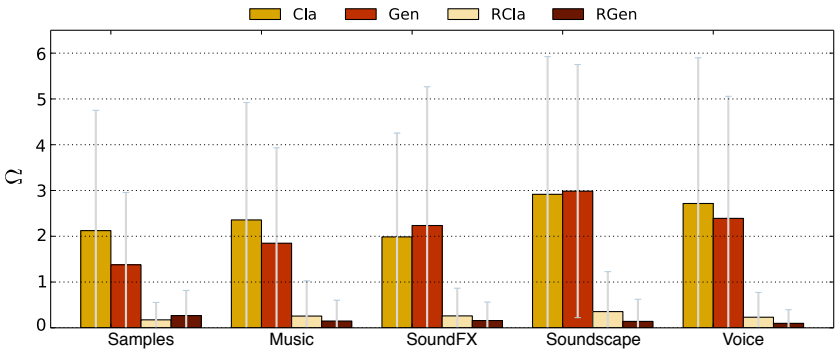


Figure 4.7: Average number of correctly predicted tags Ω per audio class and recommendation method.

English speakers tend to accept an average of 0.96 tags more than non-native ones ($p = 4.61 \cdot 10^{-3}$). Furthermore, we observe that native English speaking participants tend to annotate sounds with an average of 0.32 tags more than non-native ones ($p = 3.24 \cdot 10^{-6}$). This suggests that, in these experiments, native speakers use more tags for describing sounds than non-native speakers, and tend to accept more recommendations. 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.

4.5.2 Correctly predicted tags per audio class

To gain insight into how recommendation methods work for the different audio classes defined above (Table 4.1), we group annotated sounds by their class and recommendation method, and computed the average number of correctly predicted tags Ω for each group (Fig. 4.7). In general, sounds under SOUNDSCAPE and VOICE classes report higher Ω than sounds 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 sound in SOUNDSCAPE or VOICE classes).

It can also be observed that not all audio classes feature higher Ω for the CLA method when compared to the GEN method. SOUNDSCAPE sounds report higher Ω for GEN than for CLA, although the difference of 0.07 is not statistically significant ($p = 4.56 \cdot 10^{-1}$). SOUNDFX sounds also report higher Ω 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 larger. SAMPLE, MUSIC and VOICE classes report higher Ω for CLA recommendations, with larger Ω increases and closer to statistical significance. This suggests that the adaptation to audio categories that the CLA method performs is better exploited in VOICE, MUSIC and SAMPLE classes than in SOUNDSCAPE or SOUNDFX. We hypothesise that the vocabulary needed to accurately describe sounds from the

former classes is more reduced than the vocabulary needed for other sounds. Therefore, the class-based method can easily adapt to the class context and produce better recommendations. These recommendations probably include tags which have a narrower semantic meaning than the tags recommended with the general method. On the other hand, sounds 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 ($\mathcal{S}_{\mathbf{C}_h}$, Table 4.3). VOICE, MUSIC and SAMPLE produce smaller similarity matrices, with less tags in the vocabulary, than SOUNDSCAPE and SOUNDFX.

4.5.3 Correlation between number of uploaded sounds and the number of correctly predicted tags

All participants in our experiment were Freesound users. However, not all of them had experience in uploading and tagging sounds in Freesound. In order to get some insight in how being used to tagging sounds affects Ω, we compute the correlation between the number of uploaded sounds and the number of accepted tags, grouping sounds into the four evaluated recommendation methods (Table 4.5). To measure that correlation, we employ the Spearman's rank correlation coefficient (Corder & Foreman, 2009), with ϱ denoting the correlation coefficient and p the p-value associated with it.

We find the strongest correlation for the CLA method ($\varrho = 0.276$, $p < 3.76 \cdot 10^{-7}$). Thus, in this case, Ω 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 processes, the more recommended tags they tend to accept, especially 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. 4.5.1). 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.

4.5.4 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. In our experiments, we measured the average

Num. of uploaded sounds	CLA	GEN	RCLA	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 4.5: Average number of correctly predicted tags Ω per number of uploaded sounds and recommendation method. The ranges in the number of uploaded sounds are determined by the questionnaire that participants had to fill at the beginning of the experiment (Fig. 4.5).

time invested for annotating a sound and observed that there exists a significant correlation between the length of the sounds and the time invested to annotate them, shorter sounds being the fastest to be annotated ($\varrho = 0.24$, $p < 5.68 \cdot 10^{-19}$). This could be expected, as shorter sounds tend to be less complex and need less time for listening to them. Consistently, sounds belonging to the SOUNDSCAPE class need an average of 15 extra seconds to be described when compared to sounds belonging to other classes ($p < 8.12 \cdot 10^{-3}$). On the other hand, SAMPLE sounds need less time than the rest ($p < 3.15 \cdot 10^{-2}$). This can be explained because SOUNDSCAPE sounds are generally longer than sounds from other classes, while SAMPLE sounds tend to be shorter. Nevertheless, when comparing the four different recommendation methods, we have not observed any statistically significant differences in the average time invested for annotating sounds. Therefore, in our particular comparison, the choice of a recommendation method does not seem to affect the time needed to annotate sounds.

4.5.5 User feedback

In the last phase of the online experiment, participants were provided the opportunity to give some feedback in the form of textual comments (Sec. 4.4). Looking at these comments, 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 sounds (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 fact, approximately 29% of all tag annotations performed during the experiment were suggested by the recommendation systems[32] (i.e., were correctly predicted), but the other tags were created by users.

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

A lot of participants point out that annotation is especially hard when the sound being described is not recorded/created by the person annotating it (which was always the case in our experiment). In these 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 sound 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 sounds 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. These issues are particularly relevant in our context, where participants had to annotate sounds not created by themselves. Finally, another repeated comment is that tag suggestions are more useful for "nature" and "human-related" sounds, whereas "abstract" and "synthetic" sounds require more tags to be manually introduced before some meaningful suggestions are made. These comments are somehow aligned with the results reported in Fig. 4.7, where we see that SOUNDSCAPE and VOICE classes are the ones that report higher Ω.

4.5.6 Tag analysis

Here we have a closer look at the experiment logs in order to get some insight into the type of tags that are recommended and in which cases these are correctly predicted. We observe several interesting patterns that we believe also help comprehend in more detail tag recommendation processes in general. First of all, there are some tags which are recommended and accepted many 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 in providing some kind of general categorization to annotated sounds, but sounds only tagged with these kind of tags do clearly lack specificity in the annotations. We observe that another recommendation pattern consists of 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. This is also the case of tags like `soundscape`, `percussion-loop`, `drum-loop` or `natural-reverb`, which can typically be represented with alternative tags, compound-tags or pairs of tags such as `field-recording` (instead of `soundscape`), `loop`, `percussion`, `drum`, `natural` or `reverb`.

We also observe that there are some tags whose low acceptance can be explained 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 sounds (e.g., `multi-sample`, `improvised`). Obviously there are also some suggested tags which are not accepted because they are simply not appropriate for the sounds being described. This is the case of tags like `piano`, `guitar` or `pad`, which are sometimes recommended to sounds that clearly do not contain piano, guitar or pad-like sounds. Finally, we ob-

serve 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 that can be used as a broad categorisations of the sounds. However, recommendations are not as useful when they refer to more detailed aspects of the sounds being annotated.

4.6 Complementary results and evaluation of the tag recommendation methods

In order to complement the user-based evaluation, we also consider a systematic assessment of the different tag recommendation methods (Cla, Gen, RCla and RGen) following the methodology we described in Chapter 3. This complementary assessment follows a setup based on a tag prediction task which we now describe.

4.6.1 Methodology

For this evaluation we consider sounds and annotations of the same Freesound dataset described in Sec. 4.4. The process we follow is very similar to that described in Chapter 3 (Sec. 3.3.2). However, for each fold of the 10-fold cross-validation, we now have to follow two extra steps to set up the system for producing tag recommendations. The first step consists of training a classifier that allows the classification of the input tags into one of the five defined audio classes. We train the classifier as described in Sec. 4.2.2, but feeding the classifier only with these sounds that are present both in the training set of the current fold and in the ground truth we built when designing the system (i.e., we only use sounds from the training set that we know to which audio category they belong to).

The second step of the training phase consists of building the general tag-tag similarity matrix $\mathcal{S}$ and the matrices $\mathcal{S}_{\mathbf{C}_h}$ for every class $\mathbf{C}_h$. For this we use information from all the sounds in the training set. Notice that building $\mathcal{S}_{\mathbf{C}_h}$ requires the classification of all sounds of the training set into one of the five defined categories (Sec. 4.2.2). We perform that classification using the same classifier trained in the first step of the training phase. Hence, this classifier is not only used during the recommendation process to automatically detect the audio class of a set of input tags $\mathbf{T}_\mathrm{I}$, but it is also used to build the different tag-tag similarity-matrices $\mathcal{S}_{\mathbf{C}_h}$ corresponding to each audio class.

Similarly to Sec. 4.2.2, after the training phase we pick every sound in the evaluation set and randomly delete a set of tags $\mathbf{T}_\mathrm{D}$ from its originally assigned

tags, yielding $\mathbf{T}_\mathrm{I}$, the input to our recommendation system. The number of tags we delete is chosen uniformly at random, with only the constraint of leaving a minimum number of input tags of $|\mathbf{T}_\mathrm{I}| \geq 3$ so that there is presumably enough information for the recommender systems to provide good recommendations (see Sec. 3.3.1). This constraint also implies that, in order to be able to remove at least one tag for each sound ($|\mathbf{T}_\mathrm{D}| \geq 1$), we can only consider for evaluation the sounds that have at least four tags[33]. After we remove some tags, we run the four tag recommendation methods using $\mathbf{T}_\mathrm{I}$ as input and the similarity matrices we computed in the training phase. As evaluation measures we compute standard precision, recall, and f-measure (P, R, and F, respectively) for each evaluated sound (Eq. 2.1). Global P, R and F measures for each tag recommendation method are calculated by averaging precision, recall and f-measure across all sounds evaluated with the chosen recommendation method. Because of the nature of the tag prediction task, and as mentioned in Secs. 2.3.4, 3.3.2 and 3.5, tag recommendations in this evaluation are only considered as being "correct" recommendations if they contain tags originally assigned by the authors of the sounds. As a result, tags that could be subjectively considered as good recommendations for a particular sound but are not present in the original annotations do not count as correct predictions. Hence, the results provided by this evaluation are considered to be an underestimate of the real performance of the system.

4.6.2 Results

Results for the four evaluated tag recommendation methods (Table 4.6) are very similar to those observed in the user study (Table 4.4). 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 fewer tags than the pool in RGEN (Sec. 4.2.2). 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 $\mathcal{S}_{\mathbf{C}_h}$. We also observe that CLA and GEN feature a very similar number of recommended tags $|\mathbf{T}_\mathrm{R}|$, with an average of 3.99 and 3.88 tags, respectively.

If we analyse F as a function of the number of input tags $|\mathbf{T}_\mathrm{I}|$ and the number

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

Method	Precision	Recall	F-measure
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)

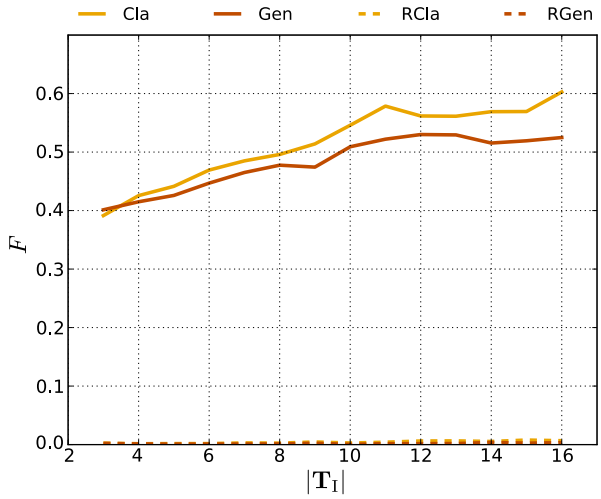
Table 4.6: Average precision P, recall R and f-measure F (standard deviation in parenthesis) for the prediction-based evaluation approach. Results are sorted by f-measure.

of recommended tags $|\mathbf{T}_\mathrm{R}|$, we can get more insight on the behaviour of the considered recommendation methods (Fig. 4.8). For instance, we see that both CLA and GEN have a tendency of increasing F as the number of input tags also increases (Fig. 4.8a). Note that, as we have seen in Sec. 4.3.2, the accuracy of the classifier also increases for larger $|\mathbf{T}_\mathrm{I}|$, which is consistent with these results. Overall, we see that the recommendation system is able to provide better recommendations when it is fed with more input tags. The opposite happens with the number of recommended tags (Fig. 4.8b). 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 ($|\mathbf{T}_\mathrm{R}| > 8$, Fig. 4.8b). 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.

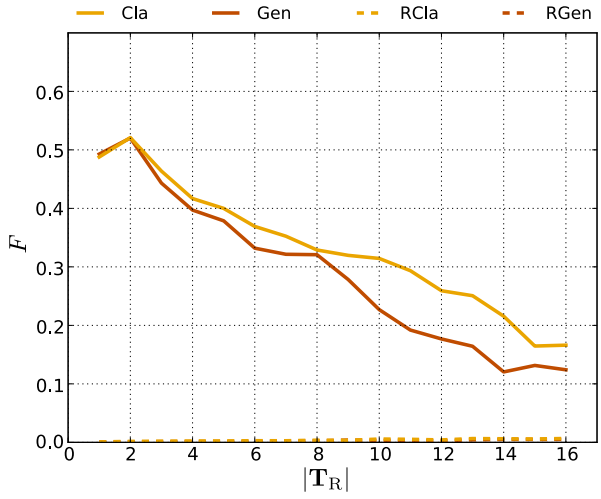
4.7 Conclusion and discussion

In this chapter we have described an extension of the best performing tag recommendation method described in Chapter 3, and thoroughly evaluated it. The method we described here extends the former in two main aspects: it automatically determines to which class a sound 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 a number of classes of resources in the particular domain, and the construction of a ground truth to train the classifier needed to perform recommendations.

We performed a user-based evaluation through an online experiment. In it, participants had to annotate several sounds 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



(a)



(b)

Figure 4.8: Average f-measure F as a function of the number of input tags $|\mathbf{T}_\mathrm{I}|$ (a) and the number of recommended tags $|\mathbf{T}_\mathrm{R}|$ (b).

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.

In general, 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 the general method, 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 VOICE 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 sound) and because tag recommenders tend to produce less meaningful suggestions in these cases.

Summarising, here we continued with the research described in the previous chapter by proposing an extended tag recommendation method that incorporates basic knowledge of the audio domain, and by comparing, with an online experiment, the extended method with the best scoring method of the previous tag recommendation scheme. In the next chapter (Chapter 5), we further investigate on tag recommendation by analysing the impact of the class-based method in a large-scale experiment on the real-world tagging system of Freesound.

CHAPTER 5

Impact of a tag recommendation system

5.1 Introduction

In the previous chapters we have described a number of tag recommendation methods and evaluated them from different perspectives. In this chapter, we perform a large-scale experiment in which we analyse the impact of a tag recommendation system in the real-world folksonomy of Freesound. More specifically, we introduce the best performing recommendation method described in Chapter 3 (RankP@α, with$\alpha = 0.15$) to the tagging system of Freesound with the classification extension described in Chapter 4, and analyse its impact on the site.

As we have seen in the literature review, many authors have hypothesised about the potential impact of tag recommendation in the folksonomies of online sharing sites (Sec. 2.3.5). Taking this into consideration, we can summarise the expected impact into the following three hypotheses:

1. *Vocabulary convergence.* A tag recommendation system should contribute to the convergence and consolidation of a shared vocabulary across the users of a tagging system (Golder & Huberman, 2006; Marlow et al., 2006; Jäschke et al., 2007; Sood et al., 2007; Zangerle et al., 2011).

2. *Quality of annotations.* A tag recommendation system should improve the quality of resource annotations in an online sharing platform (Naaman & Nair, 2008; Jäschke et al., 2012; Wang et al., 2012).

3. *Cost of the annotation process.* A tag recommendation system should reduce the cost of tagging, changing from a tag generation process to a tag recognition process (Sood et al., 2007; Jäschke et al., 2007; Wang et al., 2012).

Although there seems to be a consensus on the hypotheses, we are not aware of any study performing a deep analysis of the impact of a tag recommendation system into a real-world and large-scale folksonomy. Furthermore, even though several studies have focused on analysing aspects such as tagging behaviour or vocabulary convergence in tagging systems (Chapter 2), there is not a clearly defined set of evaluation metrics or methodology to carry out these analyses (Farooq et al., 2007).

In this chapter, we define a series of metrics to illustrate each of the three summarised hypotheses. Then, we compute the defined metrics for an extensive period of time comprising 2.5 years of Freesound analysis data, including three months after the introduction of tag recommendation. We put a special emphasis on analysing the changes observed before and after the introduction of tag recommendation. Our results give, for the first time, empirical and quantitative evidence of the validity of some of the previous hypotheses. Specifically, our results show that tag recommendation effectively contributes to vocabulary convergence, partially contributes to an improvement of the annotation quality, but does not seem to significantly reduce the cost of the annotation process. Notice that both the definition of the metrics and the analysis of its results are relevant contributions of the present chapter.

The rest of this chapter is organised as follows. In Sec. 5.2, we briefly summarise the components of the tag recommendation system that we implemented, and describe the evaluation metrics and analysis methodology. The results for all evaluated metrics, along with discussions about their implications, are reported in Sec. 5.3. We end in Sec. 5.4 with a discussion about our findings and future directions.

5.2 Methods

5.2.1 Tag recommendation algorithm

As mentioned, the recommendation method implemented in the tagging system of Freesound corresponds to the class-based approach described in Chapter 4. In summary, it is composed of the three main steps described in Chapter 3 plus the class detection step added in Chapter 4 (see Fig. 5.1). Given a set of input tags $\mathbf{T}_{\mathrm{I}}$, the recommendation method is able to generate several lists of candidate tags $\mathbf{T}_{\mathrm{C}}^{i}$ taking advantage of a tag-tag similarity matrix $\mathcal{S}_{\mathbf{C}_h}$ which is, in turn, derived from the analysis of a folksonomy $\mathcal{F}$. A particular similarity matrix $\mathcal{S}_{\mathbf{C}_h}$ is chosen after a class detection step in which the system predicts the audio class $\mathbf{C}_h$ that better fits $\mathbf{T}_{\mathrm{I}}$. The different sets of candidates $\mathbf{T}_{\mathrm{C}}^{i}$ are then aggregated into a single set of tags with scores $\mathbf{T}_{\mathrm{A}}$. Finally, a simple heuristic is applied to decide which of the tags in $\mathbf{T}_{\mathrm{A}}$ are relevant enough to form the set of recommended tags $\mathbf{T}_{\mathrm{R}}$ outputted by the recommendation method.

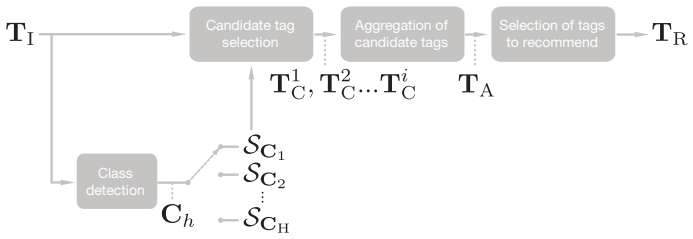


Figure 5.1: Block diagram of the tag recommendation method implemented in the tagging system of Freesound.

Figure 5.2: Screenshot of the tagging interface of Freesound after introducing the tag recommendation method. The previous interface (before the introduction of the tag recommendation), was exactly the same without the list of tag suggestions at the bottom.

5.2.2 Tag recommendation interface

Fig. 5.2 shows a screenshot of the interface for the tag recommendation system implemented in Freesound. In it, we can see the set of input tags $\mathbf{T}_\mathrm{I}$ ={`river`, `water`} and the set of recommended tags $\mathbf{T}_\mathrm{R}$ ={`stream`, `creek`, `brook`, `flow`, `liquid`, `waterfall`, `trickle`}. The list of suggested tags appears at the bottom of the text area that users use to type their tags, and it is automatically refreshed each time that users type a new tag (i.e., every time that there is a change in $\mathbf{T}_\mathrm{I}$). This means that during the annotation process of a particular sound, several lists of recommended tags are presented to the user. To introduce tags from the list of recommendations, users can either click on the elements of the list or type them manually (as they would do to introduce other tags that are not in the list). When manually typing tags, no autocomplete functionality is provided.

5.2.3 Analysis metrics

To assess the impact that the tag recommendation system has on the folksonomy of Freesound, we define a series of metrics which are meant to illustrate

Hypothesis	Metric	Expectation
Vocabulary convergence	Percentage of new tags	Decrease
	Average user vocabulary size	Increase
	User vocabulary sharing	Increase
	Sound vocabulary sharing	Increase
Quality of annotations	Average tagline length	Increase
	Percentage of misspelled tag applications	Decrease
	Tag frequency distribution	Even (see caption)
	Subjective annotation quality	Increase
Cost of the annotation process	Average tag application time	Decrease
	Average number of correctly predicted tags	Similar to Sec. 4.5.1

Table 5.1: Proposed metrics and expected observations to evaluate the hypotheses. In the case the tag frequency distribution, we expect a more even distribution across the frequency range after the introduction of tag recommendation.

the three hypotheses summarised in the introductory section of this chapter (Sec. 5.1). Rather than the observation of a single metric being affected after the introduction of the tag recommendation system, we believe the relevance of the analysis particularly remains on the observation of changes simultaneously happening in several metrics. Hence, we illustrate each hypothesis with more than one metric. Table 5.1 shows a list of the defined metrics, along with the changes we expect to observe when comparing data before and after the introduction of the tag recommendation system. Formal metric definitions subsequently follow, grouped by hypothesis.

Vocabulary convergence

- *Percentage of new tags*: This metric represents the percentage of the tag applications that were performed during a given day of our analysis period, and that had never been used before in the folksonomy (i.e., tag applications that introduce previously nonexistent tags in the folksonomy). Thus, this metric is computed on a daily basis (see Sec. 5.2.4). Considering the folksonomy model defined in Sec. 3.2.1, the percentage of new tags can be defined as

$$\Theta(n) = 100 \cdot \frac{|\mathbf{T}^n_{\text{NEW}}|}{|\mathbf{E}^n|}, \tag{5.1}$$

  where $\mathbf{T}^n_{\text{NEW}}$ is the set of tags that appeared for the first time in the n-th day of our analysis data, and $\mathbf{E}^n$ is the set of all tag applications

performed during that same day. Note that $\mathbf{T}^n_{\text{NEW}}$ can not contain duplicates (i.e., a particular tag can not be considered as being "new" more than once). High values of Θ indicate that many new tags are being created and that, therefore, the vocabulary is not converging to a finite set of terms. Our expectation is that Θ will decrease after the introduction of tag recommendation, as users will tend to reuse tags from the list of suggestions rather than creating new ones.

- *Average user vocabulary size*: This metric is also computed on a daily basis, and we define it as the total number of tag applications involving distinct tags that a user performed during a given day (i.e., the number of unique tags that a user assigned during a given day). Considering the folksonomy model defined in Sec. 3.2.1, the average vocabulary size can thus be expressed as

$$\Upsilon(n) = \frac{1}{|\mathbf{U}^n|} \sum_{u \in \mathbf{U}^n} |\mathbf{E}^{u,n}|, \tag{5.2}$$

where $\mathbf{E}^{u,n}$ is the set of tag applications involving distinct tags that user u has performed during the n-th day of our analysis data, and $\mathbf{U}^n$ is the set of users that performed at least one tag application during that same day. High values of Υ indicate that users employ a wide variety of tags for annotating their sounds, whereas low values indicate that users tend to employ a restricted vocabulary of tags. We believe that when using the tag recommendation system users will be exposed to a wider variety of tags than the ones that they would have initially thought of. Hence, we expect to observe an Υ increase after the introduction of tag recommendation.

- *User vocabulary sharing*: This metric quantifies to which extent users employ tags that have also been employed by other users. To analyse this aspect, we build a weighted network $\mathcal{U}$ where nodes represent users and edges represent the amount of tags shared between two users. Edge weights w between nodes i and j of $\mathcal{U}$ are normalised using standard Jaccard similarity. Thus, given an arbitrary period of time k for which a network $\mathcal{U}_k$ can be constructed, the weight between two nodes can be computed as

$$w_{ij} = \frac{\left|\mathbf{T}^{i,k} \cap \mathbf{T}^{j,k}\right|}{\left|\mathbf{T}^{i,k} \cup \mathbf{T}^{j,k}\right|}, \tag{5.3}$$

where $\mathbf{T}^{i,k}$ is the set of distinct tags that the user corresponding to the i-th node has annotated during the time period comprised in k (similarly for $\mathbf{T}^{j,k}$ and node j). In such a network, two users will be strongly connected if they use the same tags when annotating their sounds. Notice that, according to the definition above, every node in $\mathcal{U}_k$ has a self-loop,

i.e., for $i = j$ we have $w_{i,j} = 1$. Having defined $\mathcal{U}_k$, node strength (Barrat et al., 2004) acts as a basic indicator of the level of vocabulary sharing across users. The more strength the nodes have, the more tags users are sharing. Let L be the total number of nodes in $\mathcal{U}_k$, and ϑ_i be the node strength for the i-th node of $\mathcal{U}_k$ such that

$$\vartheta_i = \sum_{j=1}^{L} w_{ij}. \tag{5.4}$$

We define user vocabulary sharing Ψ_{u} as the average node strength over the network so that

$$\Psi_{\mathrm{u}}(\mathcal{U}_k) = \frac{1}{L} \sum_{i=1}^{L} \vartheta_i. \tag{5.5}$$

In our analysis, we build two networks $\mathcal{U}_k$ as defined above, one considering all the data after the introduction of tag recommendation and the other considering data from a reference time window before the introduction of tag recommendation (see below). We compare these two networks by computing the difference between user vocabulary sharing (average node strength) in both networks. We asses the statistical significance of that comparison by taking the series of node strengths of both networks (i.e., without computing the average) and using the Kolmogorov-Smirnov two-sample test (Corder & Foreman, 2009) for evaluating the null hypothesis that both node strength samples belong to the same distribution (we use a significance level of 0.01). After the introduction of tag recommendation, we expect to observe an increase in Ψ_{u}, as users will be highly exposed to the influence of tags used by other users, and therefore more links will be created in $\mathcal{U}$.

- *Sound vocabulary sharing*: Similar to the previous metric, we can also study the vocabulary sharing across sounds instead of users. In this way, sound vocabulary sharing represents the tags that sounds have in common. To analyse sound vocabulary sharing, we build a weighted network $\mathcal{R}$ where nodes represent sounds and edges represent the number of tags that are common in the two sounds linked by the edge. As in $\mathcal{U}$, edge weights are normalised using the Jaccard similarity. Hence, the weight w between nodes i and j of a network $\mathcal{R}_k$ computed with data from a time period k, can be defined as

$$w_{ij} = \frac{\left|\mathbf{T}^i \cap \mathbf{T}^j\right|}{\left|\mathbf{T}^i \cup \mathbf{T}^j\right|}, \tag{5.6}$$

where $\mathbf{T}^i$ is the set of tags assigned to the sound represented by the i-th node (similarly for $\mathbf{T}^j$ and node j). Notice that, in this case, the definition of w_{ij} does not include the time period k in any of its terms.

This is because all tag applications for a given sound are done at once. Therefore, if the sound was uploaded in the time period k (and thus is represented by a node in the network $\mathcal{R}_k$), all its tag applications will have also been performed during that time period k. In $\mathcal{R}_k$, two sounds will be strongly connected if they are annotated with the same tags, and we consider node strength as a basic indicator of the vocabulary sharing across sounds. Thus, we can define sound vocabulary sharing Ψ_{r} for a network $\mathcal{R}_k$ as the average node strength over that network, and compute it in the same way as described for user vocabulary sharing.

For analysis purposes, we again build two networks with data before and after the introduction of tag recommendation. The two networks are compared in terms of their node strength following the same process described above for analysing user vocabulary sharing. After the introduction of tag recommendation, we expect to observe an increase in Ψ_{r}, as users will be highly exposed to the influence of tags used by other users. Therefore, sound annotations will include these tags and more links will be created in the network $\mathcal{R}$.

Quality of annotations

- *Average tagline length*: This metric is computed on a daily basis, and we define it as the average number of tags assigned to sounds uploaded during a given day of our analysis period. Considering the folksonomy model defined in Sec. 3.2.1, the average tagline length can be expressed as

$$\Gamma(n) = \frac{1}{|\mathbf{R}^n|} \sum_{r \in \mathbf{R}^n} |\mathbf{T}^r|, \tag{5.7}$$

where $\mathbf{T}^r$ is the set of tags assigned to a resource r and $\mathbf{R}^n$ is the set of sounds uploaded and annotated during the n-th day of our analysis. High values of Γ indicate that sounds are being annotated with many tags, with potentially more comprehensive descriptions. Our expectation for this metric is to observe an increase after the introduction of tag recommendation, as the provided list of recommendations will help users to add more tags during the annotation process. In fact, even if recommendations are not appropriate, they may serve as a guide for users, and convey which kinds of information should be annotated about the sounds being described. For instance, the recommendation system could suggest a tag like `120bpm` to a sound sample corresponding to a music loop of different tempo. However, this tag might suggest to the user that she could describe tempo information, and help in this way to generate a longer tagline. Hence, we expect Γ to be increased after the introduction of tag recommendation.

- *Percentage of misspelled tag applications*: This metric represents the percentage of tag applications that contain tags with misspellings or typographical errors and that were performed during a given day of our analysis period. Considering the folksonomy model defined in Sec. 3.2.1, the percentage of misspelled tag applications can be defined as

$$M(n) = 100 \cdot \frac{|\mathbf{E}^n_{\text{MISS}}|}{|\mathbf{E}^n|}, \tag{5.8}$$

  where $\mathbf{E}^n$ is the set of all tag applications performed during the n-th day of our analysis data, and $\mathbf{E}^n_{\text{MISS}}$ is the set of tag applications performed during that same day which involve misspelled tags. In order to estimate $\mathbf{E}^n_{\text{MISS}}$, we use a straightforward approach in which we check, for each individual tag, whether if it exists or not in an English dictionary. For that purpose we use the open-source Enchant spellchecking library, with British English and American English dictionaries[34] (similarly to Guy & Tonkin, 2006). We consider that the tags which do not appear in the English dictionary contain misspellings or typographical errors. Using such a simple approach, tags consisting of proper nouns, compound words, or tags written in other languages, are most likely considered to be misspellings. However, we assume that the presence of these kind of tags is not affected by the introduction of the tag recommendation system, and thus our defined metric is meaningful enough for comparison purposes. High values of M indicate that many of the tags assigned to sounds contain misspellings. Our expectation is that M should decrease after the introduction of tag recommendation, as users will manually type fewer tags and choose them from the list of recommendations instead.

- *Tag frequency distribution*: One useful indicator of the impact of the tag recommendation system is the observation of changes in the frequency distribution of existing tags. Intuitively, tags that are very popular (i.e., that have a high frequency) tend to correspond to broader semantic concepts, while less popular tags usually correspond to narrower ones. Looking at the tag frequency distribution we can thus have an idea of users' tagging behaviour and observe if it is influenced by the tag recommendation system. To do that, we compute the frequency of tags over a period of time k such that the frequency υ of a tag t is expressed as

$$\upsilon(t,k) = |\mathbf{E}^{t,k}|, \tag{5.9}$$

  where $\mathbf{E}^{t,k}$ is the set of all tag applications involving tag t during the time period k. We consider two time periods, one with data before the introduction of tag recommendation and the other with data after tag

[34] `http://www.abisource.com/projects/enchant/`.

recommendation, and compute the complementary cumulative distribution of tag frequencies over the two periods. These kind of plots are common within the tagging literature (Bischoff et al., 2008; Robu et al., 2009), and indicate the probability that the number of occurrences of a particular tag is above a certain level. By qualitatively comparing the resulting distribution over the two periods of time, we can have an idea of which frequency ranges are more affected by the tag recommendation system. We believe that the tag recommendation system will have a bigger impact on the tags with mid frequency of occurrence, in which the agreement in not as clear as in the tags with high frequencies. Therefore, our expectation for this metric is that we will observe a more even tag frequency distribution after the introduction of tag recommendation.

Additionally, we compare the distribution of tag frequencies before and after the introduction of tag recommendation in terms of their fit into a power law distribution. As mentioned, it has been suggested that folksonomies whose distribution of tag frequencies can be fitted by a power law exhibit mature vocabularies leading to better quality descriptions (Sec. 2.3.5). Hence, we check if we observe any difference regarding this matter after the introduction of tag recommendation. To check whether a distribution is well fitted by a power law we use the method proposed by Clauset et al. (2009)[35]. This analysis is also directly related with the hypothesis that tag recommendation should contribute to the convergence and consolidation of the vocabulary.

- *Subjective annotation quality*: We are interested in analysing whether the tag recommendation system has an impact on the quality of sound annotations. To avoid having to define an objective metric for quality, we opt for measuring quality in relative terms, by comparing the subjective quality of a set of annotations before and after the introduction of tag recommendation. To do so, we set up a small online experiment where participants were presented with pairs of sounds from Freesound along with their taglines, and had to judge which sound was, in their opinion, better annotated. Every pair of sounds consisted of one sound uploaded after the introduction of tag recommendation and another sound uploaded before that. Sounds were labeled as "Sound A" and "Sound B", without providing any links to the original sounds in Freesound and without giving any hint of which sound was uploaded before and after the introduction of tag recommendation (Fig. 5.3). For every participant, sound pairs were presented in random order, and the assignment of each sound as being "Sound A" or "Sound B", was also randomised. For every pair of sounds, participants could either answer that "Sound A" was better annotated than "Sound B", that "Sound B" was better annotated than

[35] We use the open source implementation described in Alstott et al. (2014).

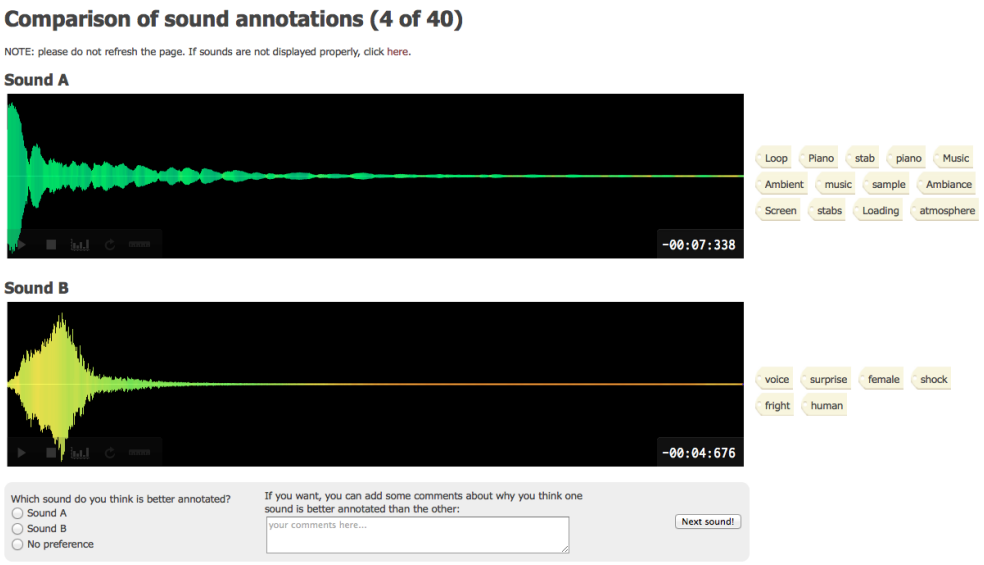


Figure 5.3: Screenshot of the online experiment interface to judge the quality of annotations.

"Sound A", or indicate that they did not think that one sound was better annotated than the other ("No preference"). If participants wanted to give further explanations for their answers, they also had the option to introduce a textual comment for every comparison.

Participants had to compare the annotation quality of a total of 40 sound pairs. To select the sounds for the experiment, we first randomly chose a set $\mathbf{X}$ of 40 sounds among those uploaded after the introduction of tag recommendation. The random selection was only constrained in such a way that all selected sounds had to be uploaded by different users. Then, we built another set $\mathbf{Y}$ of 40 sounds uploaded before the introduction of tag recommendation. In order to build $\mathbf{Y}$ and make it as similar as possible to $\mathbf{X}$ (i.e., containing similar kinds of recordings), we used the "similarity search" functionality of Freesound. For each sound $\mathbf{X}_i$, we retrieved a list of candidate similar sounds taking into account their acoustic properties represented by low-level audio descriptors[36]. Then, we pruned the lists of candidates by removing those sounds that were uploaded after the introduction of tag recommendation and by not allowing to have more than one sound uploaded by the same user. Finally, for each sound $\mathbf{X}_i$, we manually listened to the remaining candidates and selected the candidate that, in our opinion, was more acoustically similar to $\mathbf{X}_i$. Having the sets $\mathbf{X}$ and $\mathbf{Y}$, we formed the final pairs of sounds used

[36]Low-level audio descriptors mainly include spectral features such as *spectral centroid* and *MFCC*. Note that the similarity search functionality does not take into account any metadata like tags or textual descriptions.

in the experiment by iteratively selecting a random sound from each set until we got the 40 pairs determined.

We asked the team of Freesound moderators[37] to participate in the experiment, and collected data from a total of seven participants (i.e., obtaining a total seven judgements for every sound pair). Considering the collected data, we assign numerical values to the i-th quality judgement q_i performed by every participant such that

$$q_i = \begin{cases} 1 & \text{if } \mathbf{X}_i \text{ is better than } \mathbf{Y}_i \\ -1 & \text{if } \mathbf{Y}_i \text{ is better than } \mathbf{X}_i \\ 0 & \text{if no preference.} \end{cases} \tag{5.10}$$

Then, qualitative annotation quality Q is computed as the average over the union of all quality judgements q_i performed by all participants in the experiment. Let $\mathbf{Q}$ be the union of all quality judgements q_i. Then

$$Q = \frac{1}{|\mathbf{Q}|} \sum_{j \in \mathbf{Q}} \mathbf{Q}_j. \tag{5.11}$$

Note that a value of Q close to 1 indicates a preference for the annotations of sounds from $\mathbf{X}$ (i.e., sounds uploaded after the introduction of tag recommendation), while a value close to -1 indicates a preference for sounds from $\mathbf{Y}$ (i.e., sounds uploaded before tag recommendation). A value close to 0 indicates no preference. Our expectation for this metric is to obtain a positive value, indicating a tendency of considering sounds uploaded after tag recommendation as being better annotated than sounds uploaded before tag recommendation. This would suggest an increase in annotation quality.

Cost of the annotation process

- *Average tag application time*: An important indicator of how difficult it is for users to annotate sounds is the observation of the time they spend annotating them (Wang et al., 2012). For that purpose, we define the average time per tag application as

$$\Phi_{\mathrm{e}}(\mathbf{A}) = \frac{1}{|\mathbf{A}|} \sum_{a \in \mathbf{A}} \frac{\lambda_a}{|\mathbf{E}^a|}, \tag{5.12}$$

where λ_a is the duration of an annotation session a (in seconds), $\mathbf{E}^a$ is the set of tag applications performed during an annotation session a, and

[37] All sounds that are uploaded to Freesound are manually moderated by a small team of people (all of them long-term Freesound users) that ensure the appropriateness of the uploaded sounds. Hence, Freesound moderators are very familiarised with Freesound content and tagging particularities.

$\mathbf{A}$ is a set of annotation sessions. Low Φ_e values indicate that users do not need much time to add a single tag, therefore it is presumably easy for them to describe sounds.

Unfortunately, Freesound did not log information about the duration of annotation sessions before the introduction of tag recommendation. Therefore, no data was available for most of the analysed time period. To overcome that issue, during a period of time that lasted two weeks between 24 March 2014 and 7 April 2014, we altered the tag recommendation system so that it only provided recommendations to half of the annotation sessions (but logged the annotation process in both cases). Therefore, our analysis of Φ_e is carried out with data gathered only during that extra analysis period. This data includes annotation sessions for 562 sounds, one half of them annotated using tag recommendation and the other half annotated without tag recommendation. Note that this new analysis period does not overlap with the period of the main analysis (see below).

We divide the annotation session data we gathered into two sets: one containing data from sessions were tag recommendations were not provided ($\mathbf{A}^-$) and the other containing data from sessions with recommendations ($\mathbf{A}^+$). Next, we compare the average Φ_e for both sets of annotation sessions and asses the statistical significance of the difference by performing the Mann-Whitney U test with a significance level of 0.01 (Corder & Foreman, 2009). Our expectation for this metric is that sessions which provided tag recommendations will exhibit lower values of Φ_e, as users will add some tags by clicking on the tag suggestions and this will make the annotation process faster.

- *Average number of correctly predicted tags*: This is the main metric that we used in the user-based evaluation carried out in the previous chapter (Sec. 4.4). It quantifies how many of the tags assigned to a sound during an annotation process were actually suggested by the recommendation system (thus correctly predicted). We follow the definition of Eq. 4.1. However, here we define it on a daily basis such that

$$\Omega(n) = \frac{1}{|\mathbf{R}^n|} \sum_{r \in \mathbf{R}^n} \left| \mathbf{T}^r \cap \left(\bigcup_{m=1}^{M} \mathbf{T}_{\mathrm{R}}^{r,m} \right) \right|, \tag{5.13}$$

where $\mathbf{T}^r$ is the set of tags assigned to sound r, $\mathbf{T}_{\mathrm{R}}^{r,m}$ is one of the sets of recommended tags that were presented to the user in the successive M recommendations during the tagging process of r, and $\mathbf{R}^n$ is the set of sounds uploaded and annotated during the n-th day of our analysis data. Note that we can not compute Ω for data before the introduction of tag recommendation.

The average number of correctly predicted tags is an indicator of the usefulness of the tag recommendation system during the annotation process. High values of Ω indicate that many of the tags that are recommended are actually used to annotate the sounds they are recommended for, and suggest that the annotation process is less costly as tags are taken from the list of suggestions. Our expectation for this metric is to obtain similar results as in the user-based evaluation of Chapter 4 (Table 4.4).

5.2.4 Analysis methodology

The impact of the tag recommendation system is analysed by looking at the evolution of the Freesound folksonomy (gathering data directly from the Freesound database) and the logs we create every time a user annotates a new sound. Our analysis comprises data between the 21 September 2011 and 28 February 2014. The tag recommendation system was introduced on 20 November 2013. The metrics defined in the previous section are either computed on a daily basis (using data from a particular day of our analysis), or over bigger periods of time (using data data gathered from several days of our analysis). To represent daily time periods, let $\mathbf{D}$ be a vector of time periods where $\mathbf{D}_n$ corresponds to the time period of the n-th day since the beginning of our analysis data. In that vector, $\mathbf{D}_0$ corresponds to the time period of the first day in our analysis data (21 September 2011), and $\mathbf{D}_N$ corresponds to the time period of the last day for which we have analysis data (28 February 2014).

In addition to what precedes, to represent larger periods of time, we define a series of analysis windows which include data from several days of our analysis. On the one hand, let W_I be our analysis window of interest, which represents a time period including all the data after the introduction of tag recommendation (i.e., a total of 100 days from 20 November 2013 to 28 February 2014). On the other hand, let $\mathbf{W}$ be a vector of reference analysis windows where each element $\mathbf{W}_m$ corresponds to a time period of the same length as W_I (100 days), drawn from data before the introduction of tag recommendation. The window $\mathbf{W}_0$ corresponds to the last 100 days before the introduction of tag recommendation (from the 12 August 2013, to 19 November 2013), and the m-th analysis window corresponds to a time period shifted backwards in time $50m$ days. Figure 5.4 shows a graphical representation of $\mathbf{D}$ and $\mathbf{W}$, and the analysis window of interest W_I. Notice that W_I, as well as each element of $\mathbf{W}$, includes a particular range of $\mathbf{D}$ time periods (e.g., W_I corresponds to $\mathbf{D}_{N-100:N}$).

As mentioned, we are interested in comparing the results of the defined metrics for time periods *before* and *after* the introduction of tag recommendation. In the case of metrics that are computed on a daily basis, we perform the comparison by computing the average of each metric over the range of days in $\mathbf{D}$ included in the window of interest W_I and in each reference window $\mathbf{W}_m$.

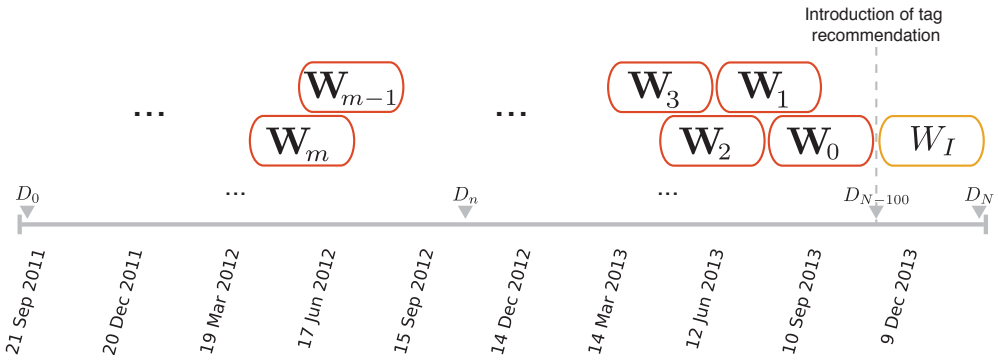


Figure 5.4: Time period vectors $\mathbf{D}$ and $\mathbf{W}$, and the analysis window of interest W_I.

Then, the average obtained from W_I is compared with the average obtained for each time period $\mathbf{W}_m$. This results in a total of M comparisons per metric. In our results section, and unless stated otherwise, we always report the results of the comparison between W_I and the $\mathbf{W}_m$ that yields the minimum difference. Hence, our results only show the case in which the tag recommendation system has the least impact. For each one of these comparisons, we assess statistical significance by taking the daily results of the metric corresponding to the compared time periods W_I and $\mathbf{W}_m$ and performing the Mann-Whitney U test with a significance level of 0.01. For the case of metrics that are not computed on a daily basis, we follow different approaches for comparing and assessing statistical significance. These approaches have been described for every particular metric in corresponding subsections of Sec. 5.2.3.

Our analysis data includes annotations for sounds of very different natures and from users with very different levels of expertise. During the analysis period, some users uploaded only one sound, while others uploaded thousands, with the average being on 12.7 uploaded sounds per user. A final point to note is that, although we do not perform any cleaning of the considered Freesound data, we remove from our consideration all tag applications performed by a specific user that, during a narrow time period within W_I (from 17 January 2014 to 27 January 2014), intensively uploaded and annotated sounds using three times more tags per sound than the average. We considered this user as being a clear outlier that could potentially bias the results of our analysis magnifying the observed impact of tag recommendation.

5.3 Results and discussion

5.3.1 Vocabulary convergence

Percentage of new tags

Fig. 5.5 shows the evolution of the percentage of new tags Θ over the considered time period. We see that, as expected, it qualitatively decreases after

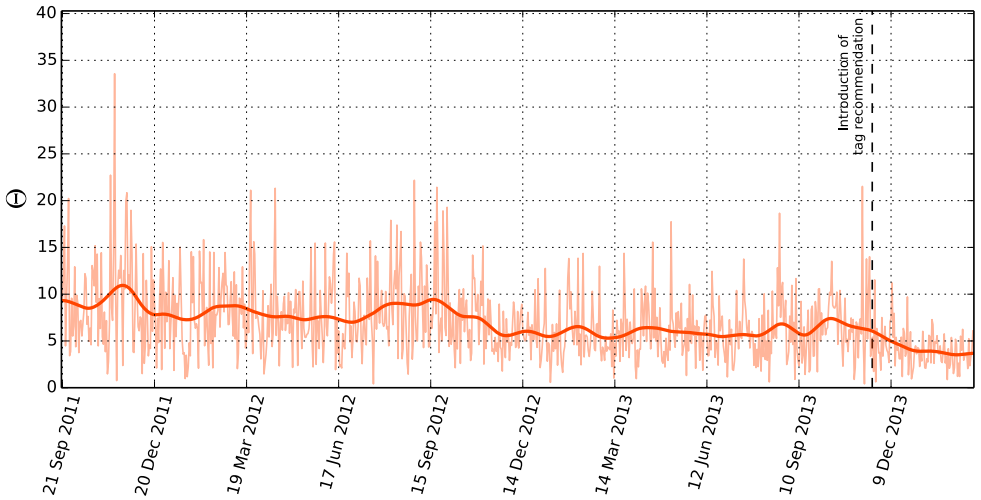


Figure 5.5: Evolution of the percentage of new tags Θ. The thinner line corresponds to computed Θ. The bold line corresponds to a smoothed version of Θ. Smoothing is performed by convolution over a moving Hann window of 51 days. That particular number of days has been arbitrarily chosen to generate an informative yet visually appealing figure. Unless stated otherwise, the same smoothing strategy is applied in the other figures in this chapter.

the introduction of tag recommendation. The minimum difference we observe between W_I and all $\mathbf{W}_m$ is a decrease of 1.7%, which is found to be statistically significant ($p = 4.01 \cdot 10^{-6}$). The maximum difference we observe is a decrease of 5% ($p = 1.26 \cdot 10^{-15}$).

The depicted evolution suggests an influence of the tag recommendation system on the percentage of new tags. However, looking at Fig. 5.5, a decreasing global trend can be qualitatively observed, even before the introduction of tag recommendation. To compensate for the existence of such a trend, we perform an extra analysis in which we apply a correction to the Θ data points obtained from W_I. The correction consists in computing a linear regression with all data points before the introduction of tag recommendation and then subtracting the linear projection of that trend to the data after the introduction of tag recommendation. Once we apply the correction to Θ over the window W_I, we repeat the comparisons with all reference windows $\mathbf{W}_m$ and observe, this time, a minimum Θ decrease of 1.5% which still remains statistically significant ($p = 5.68 \cdot 10^{-5}$). The observed global decreasing trend might be explained by a vocabulary consolidation process inherent to the tagging system, which is later accelerated with the introduction of tag recommendation.

It could be further argued that during the time period between 15 September 2012 and 14 December 2012 a localised decreasing pattern can also be observed with a similar strength to the one we observe after the introduction of tag recommendation. This decreasing pattern might be explained by the apparent

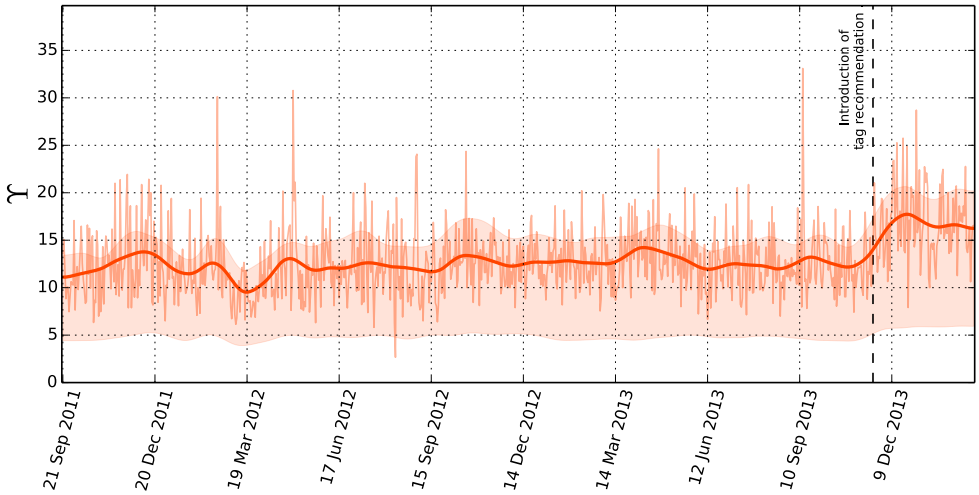


Figure 5.6: Evolution of average user vocabulary size Υ. The thinner line corresponds to computed Υ. The bold line corresponds to a smoothed version of Υ. The filled area shows the range between the lower and upper quartiles of the original data.

local increase that can be observed in the previous months, which might be provoked by a particular user uploading a significant number of sounds with many new tags. Importantly, no relevant patterns can be observed in the other studied metrics during that particular period of time (see below). Moreover, just by simple observation of Fig. 5.5, it can be spotted that the variance of Θ is smaller after the introduction of tag recommendation, thus giving more relevance to the observed decreasing pattern during W_I. As mentioned, it is the consideration of similar results from several different metrics that allows us to draw conclusions regarding the formulated hypotheses.

Average user vocabulary size

Fig. 5.6 shows the evolution of the average user vocabulary size Υ. In it, a clear impact of the tag recommendation system can be observed, as Υ consistently increases after the introduction of tag recommendation. When comparing results for the analysis window W_I and the other reference windows $\mathbf{W}_m$, we found a minimum Υ increase of 3.46 tags per user ($p = 2.303 \cdot 10^{-11}$). This demonstrates that, after the introduction of tag recommendation, users tend to use a wider variety of tags as their vocabulary size is significantly increased.

User vocabulary sharing

As described in Sec. 5.2.3, to analyse user vocabulary sharing (Ψ_u) we built two networks using data before and after the introduction of tag recommendation. In particular, we use data from the analysis windows $\mathbf{W}_0$ and W_I, respectively. The resulting network built with data from $\mathbf{W}_0$ has a total of 1,148 nodes

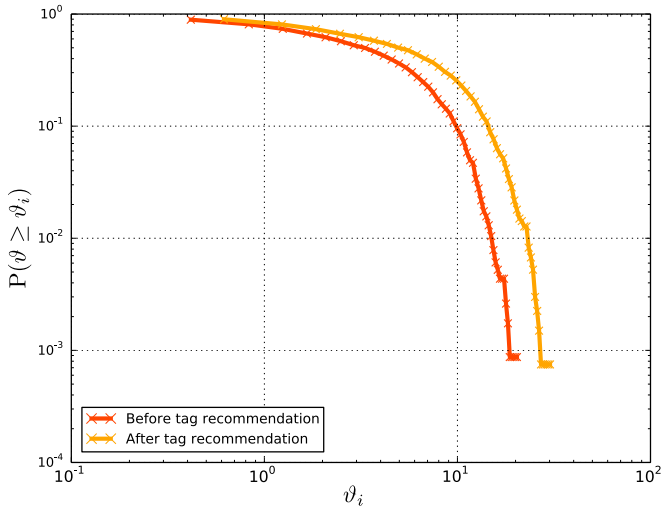


Figure 5.7: Complementary cumulative node strength ϑ distribution of user-user network $\mathcal{U}$ before and after the introduction of tag recommendation. Networks are build with data from analysis windows $\mathbf{W}_0$ and W_I respectively.

(i.e., users) and 73,240 edges (yielding a ratio of 63.79 edges per node), whereas the network built with data from W_I features 1,335 nodes and 122,474 edges (91.74 edges per node). Just by looking at these numbers, it can already be seen that users in the W_I network are much more connected among them. Fig. 5.7 shows the complementary cumulative node strength distribution of the two networks. The distribution shows that, for a given probability, the network after the introduction of tag recommendation features nodes with a higher strength. Comparing the two distributions yields a statistically significant Ψ_u increase of 2.12 ($p = 8.652 \cdot 10^{-17}$). These observations highlight that the tag recommendation system effectively favours tags sharing among users.

Sound vocabulary sharing

The analysis of sound vocabulary sharing Ψ_r reports similar results to those of user vocabulary sharing. The resulting network built with data from $\mathbf{W}_0$ has a total of 9,898 nodes (i.e., sounds) and 3,414,449 edges (yielding a ratio of 344.97 edges per node), whereas the network built with data from W_I features 12,946 nodes and 7,405,037 edges (571.99 edges per node). Again, it can already be observed that the network after tag recommendation is much more connected. Fig. 5.8 shows the complementary cumulative node strength distribution of the two networks. In this case, we also observe an statistically significant overall increase of node strengths after the introduction of tag recommendation. Interestingly, this is somewhat more relevant in the range of sounds that used to be less connected in the network (roughly for $\vartheta < 200$). The average Ψ_r increase is of 34.26 ($p = 2.606 \cdot 10^{-231}$). This result is consistent with what we find in the case of user vocabulary sharing.

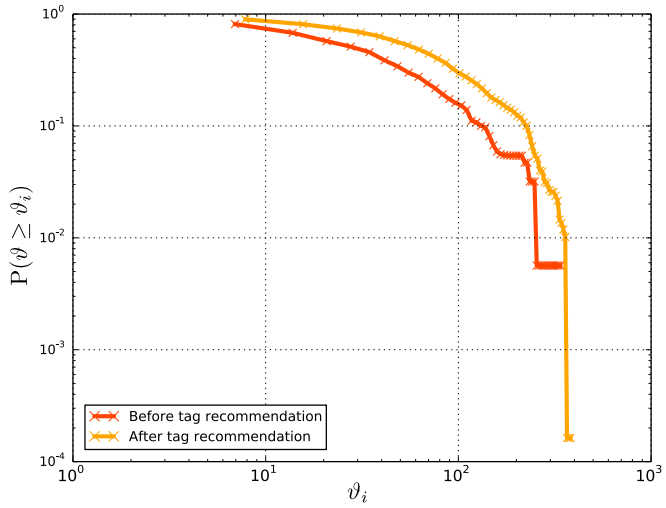


Figure 5.8: Complementary cumulative node strength ϑ distribution of sound-sound network $\mathcal{R}$ before and after the introduction of tag recommendation. Networks are build with data from analysis windows $\mathbf{W}_0$ and W_I respectively.

Discussion

We have seen that the tag recommendation system lessens the invention of new tags and that, at the same time, it increases the size of users' vocabulary and the number of tags that are shared among users and sounds. Thus, we can conclude that all users annotating sounds receive a common influence that positively affects the convergence of the vocabulary in the folksonomy by leveraging the reuse of tags, reducing the generation of new ones, and increasing the number of distinct tags in users' personal vocabulary.

We have also found that both user and sound vocabulary sharing are increased after the introduction of tag recommendation. This observation, combined with the increase in users' vocabulary size, leverages the value of sound annotations. It reveals a better agreement on the vocabulary of tags used to annotate sounds and also an increase of its size. Therefore, sounds are described using a more coherent and complete vocabulary.

5.3.2 Quality of annotations

Average tagline length

Fig. 5.9 shows the evolution of the average tagline length Γ. We observe a clear increase after the introduction of tag recommendation. Comparing results for the analysis window W_I and reference windows $\mathbf{W}_m$, we observe a minimum Γ increase of 1.32 tags per sound ($p = 7.553 \cdot 10^{-6}$). Similarly to what we noted in Sec. 5.3.1, Fig. 5.9 seems to show a global increasing tendency already before the introduction of tag recommendation. We repeated

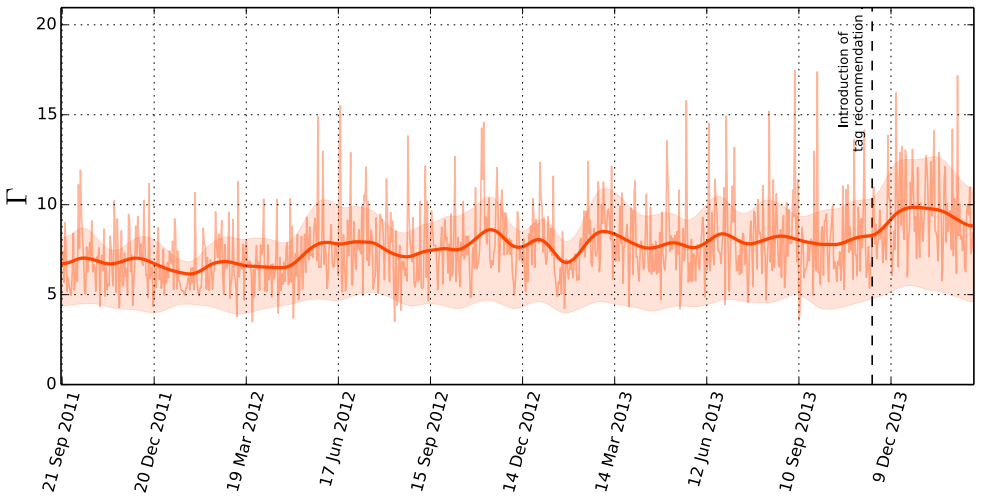


Figure 5.9: Evolution of average tagline length Γ. The thinner line corresponds to computed Γ. The bold line corresponds to a smoothed version of Γ. Filled area shows the range between the lower and upper quartiles of the original data.

the same extra analysis of that section (i.e., computing the linear regression of data before the introduction of tag recommendation and correcting Γ in W_I with the linear projection of the trend) and still observed a statistically significant minimum Γ increase of 1.22 tags per sound ($p = 3.65 \cdot 10^{-5}$). Considering the average tagline length for the time periods before and after the introduction of tag recommendation, the observed increase means that sounds are annotated with approximately 20% more tags when users are influenced by the tag recommendation system. This observation is also supported by looking at the histogram of tagline lengths before and after the introduction of tag recommendation (Fig. 5.10). The increase of the average tagline length suggests that annotations performed using the recommendation system are more comprehensive and, presumably, of better quality than annotations performed without the recommendation system.

Percentage of misspelled tag applications

Fig. 5.11 shows the evolution of misspelled tag applications M. As expected, we observe a slight decreasing tendency in M after the introduction of tag recommendation When comparing results for the analysis window W_I and the other reference windows $\mathbf{W}_m$, we find a minimum M decrease of 1.4% (not statistically significant), and a maximum decrease of 5% (statistically significant, with $p = 4.775 \cdot 10^{-5}$). Hence, this shows that the introduction of tag recommendation has a moderate impact on misspelled tags, helping users to generate up to 5% less tags with misspellings.

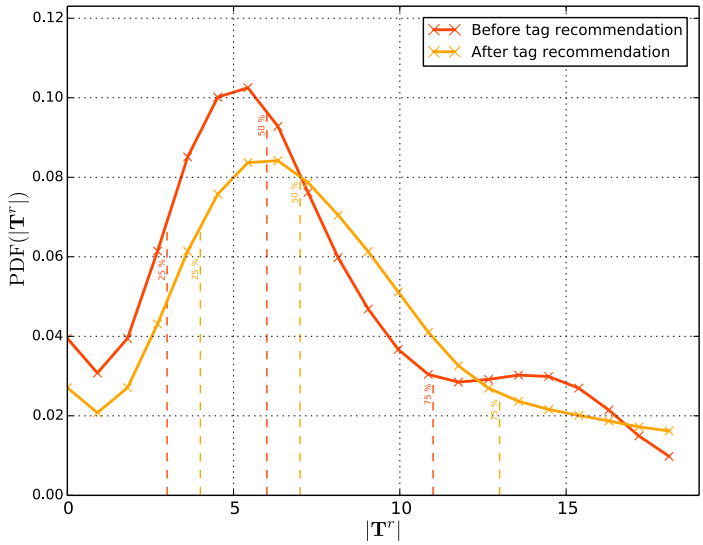


Figure 5.10: Probability density function of tagline lengths $|\mathbf{T}^r|$ before and after the introduction of tag recommendation. Data is drawn from the analysis windows $\mathbf{W}_0$ and W_I, respectively. Smoothing is performed using an arbitrarily chosen Hann window of 11 points. Dashed vertical lines with attached percentage values indicate the percentage of sounds whose tagline length is less than or equal to what is indicated in the corresponding line position.

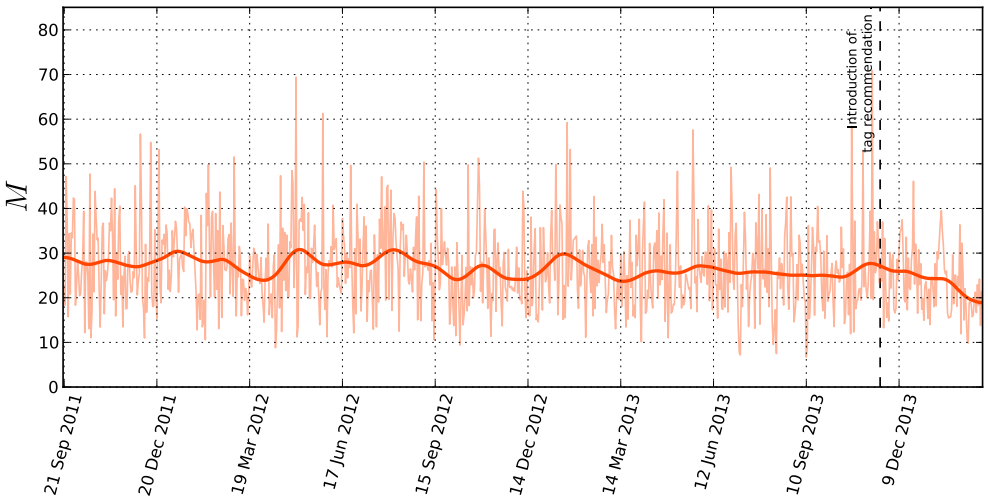


Figure 5.11: Evolution of the percentage of misspelled tag applications M. Similarly to the previous figures, the thinner line corresponds to computed M. The bold line corresponds to a smoothed version of M.

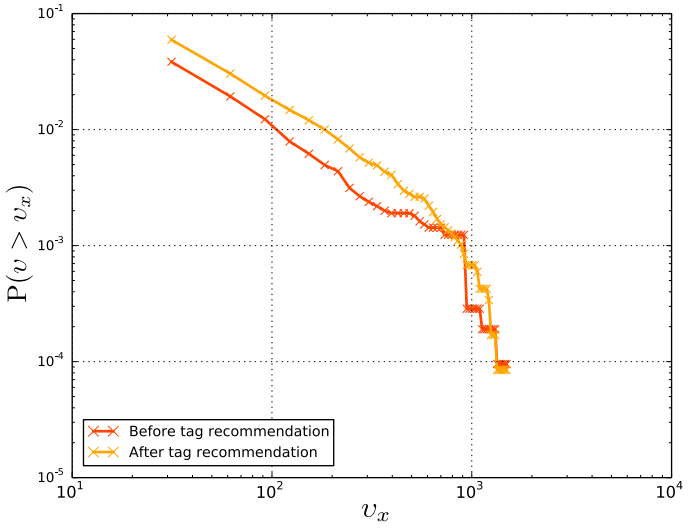


Figure 5.12: Complementary cumulative tag frequency v distribution before and after the introduction of tag recommendation. Data is drawn from the analysis windows $\mathbf{W}_0$ and W_I, respectively.

Tag frequency distribution

Fig. 5.12 shows the complementary cumulative tag frequency distribution before and after the introduction of tag recommendation. It can be observed that the distribution after the introduction of tag recommendation tends to be more even, particularly reinforcing the usage of tags in the low and mid frequency ranges (tags with less than 800 occurrences). This means that less popular tags gain importance after the introduction of tag recommendation. Less popular tags typically correspond to narrower semantic concepts, which are used to bring more details to sound annotations. Again, this observation is consistent with previous observations regarding vocabulary convergence. It reflects an increase in both user and sound vocabulary sharing, as tags with less frequency gain importance and start being more widely used. It also suggests that annotations after the introduction of tag recommendation are more detailed as the usage of tags in the low and mid frequency ranges is reinforced.

To complement these results, we evaluated how well tag frequency distributions corresponding for the analysis windows $\mathbf{W}_0$ and W_I fit into a power law distribution. In both cases, the analysis shows a better fit for a log-normal distribution rather than a power law distribution. However, the tag frequency distribution after the introduction of tag recommendation shows a better fit for the power law than the distribution before tag recommendation, which may also suggest the presence of a better converging vocabulary yielding more meaningful descriptions (Sec. 5.2.3).

Subjective annotation quality

We analyse the results of the online experiment described in Sec. 5.2.3 and observe a subjective annotation quality of $Q = 0.075$ (0.81 standard deviation). One third of the quality judgements performed by the participants correspond to "No preference" judgements ($\mathbf{Q}_j = 0$). If we discard these judgements, the subjective annotation quality is increased to $Q = 0.114$ (0.99 standard deviation), meaning that in 55% of the judgements the sounds described using the tag recommendation system are considered to be better annotated. These results indicate that participants in the experiment have a slight tendency to consider annotations of sounds described using the tag recommendation system as being better than annotations of sounds made without the tag recommendation system. To further validate these results, we computed Cohen's kappa coefficient to measure the agreement among the quality judgements performed by the participants in the experiment (Carletta, 1996). After all possible pairwise comparisons between the different participants in the experiment, we observe an average kappa coefficient of 0.22. Thus, participants in the experiment tend to agree in their judgements. This reinforces our previous observations.

During the experiment, participants also provided textual comments about some of their quality judgements. In general, participants used comments to explain the reason why they considered sounds to be badly annotated. Among these reasons, the most common ones are the presence of misleading or uncompleted annotations, the presence of tags not related to the sound being annotated, and the presence of tags with typographical errors. In the participants' sample, all these reasons are reported evenly for sounds uploaded before and after the introduction of tag recommendation.

Discussion

We have seen that the average number of tags used to annotate a sound is larger after the introduction of tag recommendation. A similar observation is made in a study by Ames & Naaman (2007), in which two mobile phone applications for uploading photos to Flickr are compared. One of the applications features a tag recommendation system to aid users in the tagging process, and an increase in the average tagline length is observed for those photos uploaded with that application.

The fact that the average tagline length increases after the introduction of tag recommendation also reinforces the previously discussed observations regarding vocabulary convergence. Tag recommendation yields more tag applications and potentially more comprehensive sound annotations, and yet fewer new tags are created while vocabulary sharing is increased. Hence, our results indicate that sound annotations after the introduction of tag recommendation are done using a more coherent and complete vocabulary of tags. This fact seems to be further confirmed by the results of the online experiment we set up to analyse

qualitative annotation quality, as participants on this experiment preferred annotations of sounds uploaded after the introduction of tag recommendation.

The tag frequency distribution we observe after the introduction of tag recommendation also supports the increase in the convergence of the vocabulary. In this case, a better agreement is reached specially for those tags with lower frequencies of occurrence. Thus, we could say that there is a better agreement on the tags users choose to annotate specific concepts, which leverages the value (and thus the quality) of the annotations.

Finally, we also observed that tag recommendation helps users in slightly reducing misspellings in the tags they introduce. This also supposes an improvement in the quality of annotations. However, the impact we observe is rather limited, which may be explained by several factors. Firstly, the way in which we estimate misspelled tags is not perfectly accurate and thus some noise is present in the metric (see Sec. 5.2.3). Secondly, the nature of the tag recommendation system does not prevent itself from actually recommending tags with misspellings. Hence, even if it is intuitively less likely that misspelled tags will feature a strong similarity with any of the input tags, it is still possible that these are recommended. Finally, we can only expect tag recommendation to effectively help in reducing misspellings for the tags that are actually suggested by the system and correctly predicted. As we describe below in Sec. 5.3.3, approximately 19% of the tags of a tagline are correctly predicted, and this can be taken as a rough estimate of an upper bound for the decrease in the percentage of misspelled tag applications. Furthermore, even when relevant tags are recommended by the system and are correctly predicted, many users still prefer to manually type them instead of clicking on the list of suggestions, which may still lead to misspellings (see Sec. 5.3.3). That being said, overall results regarding the quality of annotations suggest that the introduction of tag recommendation has a moderate yet positive impact on this aspect.

5.3.3 Cost of the annotation process

Average tag application time

Fig. 5.13 shows the probability density function of the average time per tag application Φ_e with and without the use of the tag recommendation system. Even though we observe a small average decrease in Φ_e for annotation sessions using the tag recommendation system, it is found to be not statistically significant ($p = 0.83$). This means that no substantial difference on the time needed to perform a tag application can be reported. However, if we look at the total amount of time invested in annotating every sound (instead of every tag), we do observe a statistically significant average increase of roughly 35 seconds per sound after the introduction of tag recommendations ($p = 6.2 \cdot 10^{-3}$), which represents an increase of approximately 20%. This is consistent with the 20% increase of the tagline length we observed in Sec. 5.3.2. Thus, in general, we

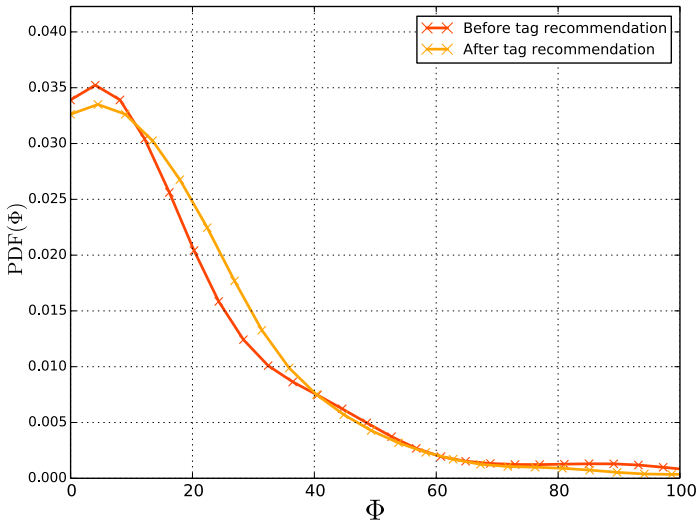


Figure 5.13: Probability density function of the average tag application time Φ_e with and without the tag recommendation system. Curves are smoothed using a Hann window of 11 points.

could say that users need at least the same amount of time to perform a single tag application as they needed before using the system. However, annotations are longer and therefore users spend more time annotating sounds.

Average number of correctly predicted tags

As explained in Sec. 5.2.3, the average number of correctly predicted tags Ω can only be computed with data drawn from W_I. Computing it on a daily basis shows that an average of 2.10 tags (from those finally assigned to sounds) were suggested by the recommendation system. This corresponds to approximately 19% of tags in a tagline. This is similar to what we found in Chapter 4 (Table 4.4), in which an average of 2.41 tags were correctly predicted by the class-based tag recommendation method (corresponding to approximately 29% of the tags in a tagline). Hence, according to these results, the tag recommendation system behaves similarly both in the real-world and in a controlled environment, with a certain tendency of users accepting fewer tags in the real world.

Among the correctly predicted tags, we make a distinction between those that are added to the tagline by users clicking on the corresponding tag in the list of suggestions, and those that are manually typed by users. If we only consider the tags that are added to the tagline by users actively clicking on the suggestion, we observe an average number of 1.58 correctly predicted tags, corresponding to approximately 13% of the tags in a tagline. This suggests that, in many occasions, users still prefer to manually type the tags instead of switching to the mouse and clicking on the list of suggestions. In general, these results show that, even though an important part of the final tagline

for a sound can be constructed using tags suggested by the recommendation system, the majority of these tags have to be generated by users themselves, and are not necessarily related with those suggested by the system.

Discussion

Contrary to what we expected, we have observed that the tag recommendation system does not seem to have a significant impact on the cost of the annotation process. Although we have seen that users need significantly more time to annotate individual sounds when using the tag recommendation system, we have also seen that this increase can be attributed to the proportional increase of the average tagline length. Hence, the actual time required for every individual tag application does not significantly change. Furthermore, we observed that most of the tags assigned to sounds are not drawn from the list of recommended tags, meaning that most of the annotation process still consists of a generation process where users create tags from scratch rather than a recognition process where users validate tags from a list of suggestions.

There are several potential reasons why we do not observe the expected impact on the cost of the annotation process. On the one hand, we observed that only 13% of the tags in taglines are added from the list of suggestions by actually clicking on them. Hence, assuming that it is faster to click on tags rather than to manually type them (which is probably not always true), the impact we can expect on the time required for introducing tags should be lower than that 13%. Also, it seems intuitively plausible that users need more time to generate the tags (or recognise them from a list) than to actually introduce them. Hence, the potential impact of lessening the time required for introducing tags is further reduced. On the other hand, the impact of the recommendation system is again limited by the fact that most of the introduced tags are not drawn from system recommendations, and thus an important part of the annotation process does not significantly change after the introduction of tag recommendation. In fact, our results might be suggesting that the cost of the recognition process is not actually lower than the cost of the generation process. This also seems reasonable, as the union of all recommended tags for a given sound is much larger than the length of the actual tagline (i.e., new tags are recommended every time that a tag is added to the tagline, see Sec. 5.2.2). Therefore, the recognition process must operate over a large set of tags.

Finally, we believe that our metrics regarding the cost of the annotation process are highly dependent on the particular interface of the recommendation system. Also, the recommendation interface can have different impacts according to how users adapt to it. Unfortunately, our analysis does not contain data to be compared coming from other recommendation interfaces. However, to get some more insight into that aspect, we repeated the calculations of the average tag application time but this time considering experienced and non-

experienced users separately. We divided users according to the number of sounds they uploaded during our analysis period. In particular, we set the threshold at the third quartile of the distribution of uploaded sounds per user, which corresponds to 7 uploaded sounds. What we observe is that the average tag application time after the introduction of tag recommendation increases for non-experienced users and decreases for experienced users by a similar amount of about 3 seconds per tag application ($p = 2.15 \cdot 10^{-3}$ and $p = 3.65 \cdot 10^{-3}$, respectively). This shows that experienced users were able to take advantage of the recommendation interface and generate annotations slightly faster, but it also shows that the interface had a negative impact on non-experienced users, apparently increasing the cost of the annotation process. This could be explained because experienced users probably have a better understanding of the tagging process and can easily interpret and take advantage of tag recommendation. Nevertheless, we think that to draw more consistent conclusions regarding the impact of tag recommendation on the cost of the annotation process, further research should be carried out.

5.4 Conclusion

In this chapter we have analysed the impact of a state of the art tag recommendation system into the real-world folksonomy of a large-scale sound sharing platform, Freesound. After a the review of current related work done in Chapter 2 (Sec. 2.3.5), we have identified three main hypotheses regarding the impact that such a system should have when introduced into a tagging system, and we have defined several reusable metrics to evaluate that impact. We have analysed data comprising of a period from 21 September 2011 to 28 February 2014, the last three months of which correspond to data after the introduction of tag recommendation. To the best of our knowledge, these kind of quantitative analyses have not been done before using large-scale data from a real-world folksonomy. Hence, no empirical assessment of the three identified hypotheses was available to date.

Our results show a significant impact of tag recommendation into most of the metrics we defined. However, the result of a single metric in isolation is probably not entirely relevant in our analysis. Instead, the fact that we observe how the changes on several metrics can be explained by some of the outlined hypotheses gives a particular value to our analysis. Overall we observe that the first hypothesis (regarding vocabulary convergence) is clearly validated, that the second one (regarding the quality of annotations) only seems to be partially validated, and the third one (regarding the cost of the annotation process) does not seem to be validated. However, we believe the latter is particularly dependent on the annotation interface, and that its impact could be greatly improved by designing an interface specifically focused on reducing the cost of the annotation process.

Although in this work we only analyse data in the context of Freesound, we believe that our results are, to some extent, indicative of the impact that tag recommendation can potentially have in other tagging systems. However, tagging systems of different nature may react differently to the introduction of a tag recommendation system. An important aspect here is to take into account the motivations that users have for tagging their resources. In narrow folksonomies such as Freesound and Flickr, users typically tag their content so that other users (and also themselves) can easily find it in the future. However, resources are only annotated once, and therefore the tags added by the uploader of a resource should also be meaningful to other users of the platform. Contrarily, in broad folksonomies such as Delicious and CiteULike, resources are tagged multiple times by several users, and thus the main motivation for tagging is users' self organisation of the content, without necessarily considering the global context of the sharing platform (Secs. 2.2.2 and 2.2.3). As a result, very different tagging styles can arise because of the particularities of these two kinds of tagging systems. The tag recommendation system that we use here is designed for narrow folksonomies. It does not try to personalise recommendations to particular users' tagging behaviours, but instead it learns from parts of the folksonomy on the basis of five audio classes (Sec. 4.2.2). Hence, we expect it to have a bigger impact in tagging systems featuring narrow folksonomies, where the more uniform across users a tagging style is, the better the platform becomes in providing content to other users.

Importantly, the metrics and analysis methodology described here are applicable to other collaborative platforms either featuring broad or narrow folksonomies. To further assess the validity of our results, an analysis with data coming from other tagging systems and tag recommendation systems should be performed. The main obstacle for carrying out this analysis is the limited availability of comprehensive tagging data, including annotations performed *with* and *without* the use of a tag recommendation system, and that comprise user activity for as long a period of time as the the one we analysed.

There are several aspects of the data we already collected that could be further researched to gain more insight into the impact of the tag recommendation system. Firstly, we do not perform any study of the generated taglines at the semantic level. By applying techniques for mapping tags to semantic concepts or categories (e.g., Cantador et al., 2011), we could analyse the impact of the recommendation system at the semantic level, and see if it effectively shapes tagging behaviour to a more extensive usage of particular kinds of tags such as content-related or self-organisational tags. Similarly, it could be further studied if other typical problems of tagging systems such as synonymy or polysemy are in fact affected by the use of a recommendation system. Secondly, in the current work we just introduced the concept of user experience when analysing our results in Sec. 5.3.3. It would be interesting to further investigate this aspect by analysing the impact of the recommendation system to other evalu-

ation metrics when considering users with different levels of expertise. Thirdly, another way in which the current analysis could be further developed would be with the use of network analysis techniques to inspect the user-user and sound-sound networks built on the basis of shared tags. Using such analysis, it would be interesting to evaluate the existence of community structure in those networks and to see how potential communities in both networks might be related. For example, we could investigate if there are strongly connected communities of users that annotate sounds with a particular tagging style, and then see how the introduction of tag recommendation would affect those communities.

In our opinion, the biggest future challenge in tag recommendation is the design of systems that have a bigger impact on the quality of annotations. Annotations are very subjective and difficult to evaluate. However, a recommendation system could be designed to particularly focus on that issue by driving recommendations at higher semantic levels, for example being able to select candidate tags for recommendation in terms of variety and coverage of different semantic facets. In order for tag recommendation systems to have a deeper impact in the tagging behaviour and in the quality of annotations, we probably need to evolve the basic tag recommendation methods into *assistive* processes where we can better guide users during the annotation process. In the following chapter (Chapter 6), we explore this perspective by proposing an extension of the current class-based tag recommendation method which takes advantage of a domain-specific ontology to drive the recommendation process.

CHAPTER 6

A new perspective: ontology-based tag recommendation

6.1 Introduction

In this chapter we present a new perspective on tag recommendation systems and explore how can it tackle some of the tagging issues that have been identified in the previous chapters. The goal is to design a tag recommendation system that further improves the quality of resource annotations. In particular, the recommendation system is focused on helping users to generate more comprehensive, coherent and semantically meaningful resource annotations.

A particularity of the folksonomies emerging from tagging systems is that tags are organised in a flat hierarchy, typically detached from a uniquely identifiable semantic meaning (Golder & Huberman, 2006; Halpin et al., 2006). Hence, tags are not restricted to a predefined set of concepts or a fixed vocabulary. We have seen that this has the advantage of enabling a certain flexibility and ease of use from the users' point of view (Sec. 2.2). This is one of the reasons why tagging systems have succeeded as a popular organisation system in online sharing platforms (Shirky, 2005; Halpin et al., 2006; Cattuto, 2006). However, we have also seen that this approach presents some disadvantages because different tagging conventions may coexist in a single folksonomy, and because the semantic meaning of tags can not be unambiguously determined (Sec. 2.2.4).

The flexibility of user-generated folksonomies is often opposed to the accurateness and rigidity of *ontologies*, which are designed by domain experts. Ontologies provide, for a given domain, an unambiguous formalisation of its concepts, entities and their relations. Hence, where folksonomies feature free-form textual labels with no predefined semantic meaning, ontologies feature detailed concept hierarchies interlinked with semantically meaningful relations.

Although folksonomies and ontologies appear to be opposed ways in which knowledge can be represented, some authors suggest that these are, in fact, complementary approaches that can be combined. For instance, some authors propose techniques for analysing folksonomies and automatically generating tag hierarchies or identifying simple semantic relations between tags (Merholz, 2004; Halpin et al., 2006; Heymann & Garcia-Molina, 2006; Mika, 2007; Hwang, 2007). These automatically derived hierarchies and semantic relations can be used to aid ontology creation processes. Other authors propose to enhance the semantic value of folksonomies by establishing unambiguous relations between tags and concepts defined in an ontology (Good et al., 2007; Passant, 2007). Similarly, other authors suggest to model folksonomies through the use of ontologies, enabling the inclusion of semantically structured content in folksonomies. In this direction, several *tagging ontologies* have been proposed which conceptualise the different agents involved in a tagging process (Newman, 2005; Limpens et al., 2009b; Echarte et al., 2007; Passant & Laublet, 2008; Kim et al., 2010; Ding et al., 2010). The use of tagging ontologies allows the definition of semantic relations between tags that can be used, for example, to tackle synonymy and ambiguity problems. Furthermore, tagging ontologies allow the interoperability of folksonomies among different sharing platforms by unifying the way in which tagging information is modelled. A comprehensive literature review of works combining folksonomies and ontologies can be found in Limpens et al. (2009a).

In this chapter, we explore the idea of combining folkosnomies and ontologies to improve the tag recommendation system described in the previous chapters. For this purpose, we define an ontology which extends a previously existing tagging ontology (see below). The ontology that we use, besides formalising tagging concepts, allows the categorisation of tags and resources into semantically meaningful categories. More specifically, it can categorise tags into a number of information facets which are particularly relevant in the audio domain. For example, the ontology specifies that tags like `guitar` and `violin` annotate musical instruments, and that tags like `english` or `german` describe the spoken language of a sound. Furthermore, the ontology defines a number of broad audio categories and relates them to the aforementioned tag categories. In this way, we are able to specify which information facets are relevant for every audio category. Following the previous example, the ontology can specify that tags describing musical instruments are typically relevant for music recordings, while tags indicating a spoken language are most relevant for voice recordings. Taking advantage of this knowledge, we propose an ontology-based tag recommendation system that is able to implement two features which clearly distinguish it from previous approaches. On the one hand, tags recommended by the system are not presented to users as a single list of suggestions, but grouped into the different information facets defined by tag categories in the ontology. On the other hand, the system can predict which information facets

are relevant for a given sound, and then suggest users to add tags that cover these facets. In this way, the tag recommendation system assists users not only by recommending tags, but also by helping them in choosing which kind of information is relevant for describing a particular sound.

Little research has been carried out on using ontologies to drive tag recommendation systems, and the followed approaches are conceptually different from the one we take here. In the works by Adrian et al. (2007) and Prokofyev et al. (2012), a tag recommendation system is described for textual resources in which natural language processing techniques are used to identify relevant keywords in a resource. Then, these keywords are matched against ontology concepts of external knowledge bases to retrieve other related concepts to be presented as tag recommendations. Hence, in these cases, the ontologies are not used to guide the recommendation process nor embed domain-specific knowledge relevant for the annotation process. Guy & Tonkin (2006) introduced an idea which is similar to that of guiding the recommendation process. In particular, they suggested that tagging can be improved by "providing users with a set of helpful heuristics that promote good tag selection, such as a checklist of questions that could be applied to the object being tagged, in order to direct the tagger to various salient characteristics", but no further research was carried out. Chen et al. (2008), describe a tag recommendation system for images in which tags are presented to users organised in a number of predefined categories. The categories defined by Chen et al. (2008) are, in fact, resource categories which group images into broad categories such as "portrait", "animal" or "architecture". Given an image to annotate, the recommendation system can estimate the most relevant categories by computing content-based similarity measures with already categorised images in a ground truth. Then, the system can recommend the most popular tags for each of the estimated relevant categories. Conversely, our approach is focused on taking advantage of the combination of tag categories and resources categories, and makes use of an ontology that formalises these categories and that allows us establish further semantic relations.

The interface of the ontology-based tag recommendation system described in this chapter allows users to introduce tags in an "attribute:value" fashion, in which the "attribute" represents a tag category and the "value" is the actual tag[38]. For example, an hypothetical attribute-tag like `instrument:guitar` has the attribute "instrument" and the value "guitar". The attribute clarifies the semantic context of the actual tag, specifying in this case that "guitar" is an "instrument". Users can select a tag category, and then the system provides specific tag recommendations tailored to that category. However, besides choosing tags from the list of recommendations, users can also type their own, therefore creating new tags for a given tag category. In this way, users are able to contex-

[38]To clarify further explanations, we will refer to those tags that are introduced with a tag category as *attribute-tags*.

tualise tags in a particular tag category, making their semantic meaning more explicit. In a sense, the concept of tag categories included in the tag recommendation system is extended to the whole tagging interface. Tag categories are therefore not only useful to guide the annotation process and provide tag recommendations, but also to allow the introduction of tags with less ambiguous semantic meaning by explicitly indicating their semantic facets (Halpin et al., 2006). This is similar to the idea of *triple-tags* introduced by Catt (2006), which was later used in the Flickr API[39] under the name of *machine-tags*. Triple-tags are normal tags formatted with a specific syntax that allows the precise specification of the meaning of a tag. By using a syntax such as "namespace:attribute=value", tags can be used, for example, to precisely specify geolocations (e.g., {`geo:lat=53.1234`, `geo:long=-2.5678`}). Hence, as far as tags contain a known namespace and attribute, their meaning can be easily interpreted. Using triple-tags, the Flickr API can respond to complex queries that operate on the namespace, attribute and value of the tags. However, to our knowledge, triple-tags can only be used through the Flickr API, and no user interfaces nor recommendation systems have been developed for them.

To evaluate the tag recommendation system described here, we perform an online experiment with more than 200 participants. We then compare the ontology-based tag recommendation system with our previous class-based recommendation system described in Chapter 4. In this online experiment, a group of participants annotate a pool of sounds using the ontology-based system, while another group annotate the same sounds using the class-based system. Then, we define a number of metrics (some of them already used in the previous chapters) to compare both systems. Furthermore, to complement the results of that experiment, we perform a second experiment in which the ontology-based interface is deployed in Freesound. With this second experiment, we collect real-world data usage of the interface that we analyse and compare with the results from the first experiment. In general, our results show that the ontology-based tag recommendation system can effectively help in improving sound annotations in those cases where users spend enough time and give enough importance to the annotation process.

The rest of the chapter is organised as follows. First, we describe in detail the ontology-based tag recommendation system, including the design and population of its ontology, and the user interface (Sec. 6.2). Then, we describe the online experiments and metrics that we used to evaluate the system (Sec. 6.3). Evaluation results are reported in Sec. 6.4, and the chapter ends with a discussion about our findings (Sec. 6.5).

[39] `http://www.flickr.com/groups/api/discuss/72157594497877875/`

6.2 Method

6.2.1 Ontology design

The ontology that we use to drive our tag recommendation system is an extension of the Modular Unified Tagging Ontology, or MUTO[40] for short (Lohmann et al., 2011). The MUTO ontology builds on top of previously existing tagging ontologies, and it was originally proposed to unify them. For this reason, we use it as a starting point for our ontology. In the core of the MUTO ontology, the `muto:Tagging` class is defined along with several object properties[41] to indicate, among others, a resource that is tagged (`muto:hasResource` of type `rdfs:Resource`), the tag assigned to the resource (`muto:hasTag` of type `muto:Tag`), and the user that made the tag assignment (`muto:hasCreator` of type `sioc:UserAccount`). Particular users, tags and resources are modelled as instances of the classes `sioc:UserAccount`, `muto:Tag` and `rdfs:Resource` respectively. Using such an ontology, it is possible to model the contents of a folksonomy in a structured manner. However, the MUTO ontology (together with the other existing tagging ontologies) is focused on the representation of the tagging process, but does not a priori incorporate other kinds of knowledge which may be specific to the particular domain of a tagging system. To overcome that limitation, we propose a simple extension of the MUTO ontology which meets the requirements of our ontology-based tag recommendation system.

We extend the tagging ontology in several ways. First, we add a number of subclasses to the `muto:Tag` class. These subclasses are used instead of `muto:Tag`, and therefore the tags in our ontology are modelled as instances of these subclasses (right side of Fig. 6.1). In our ontology, `muto:Tag` subclasses conceptualise a number of tag categories according to different kinds of information facets conveyed by tags. Hence, a tag category groups a set of tags that share some semantic meaning in the audio domain. For example, we define tag categories[42] such as `fso:InstrumentTag` or `fso:MicrophoneTag`, which include tags that convey information about the musical instruments present in a recoding, or about the microphones that were used[43]. A complete list of the different tag

[40] MUTO: Modular Unified Tagging Ontology. `http://muto.socialtagging.org/core/v1.html`.

[41] In ontologies, object properties are used to relate instances (individuals) of particular classes. For example, using object properties it can be specified that a particular instance of a class `:ClassA`, is `:similarTo` an instance of `:ClassB`. In that case, `:similarTo` is an object property with a particular semantic meaning that must be defined in the ontology. Object properties can impose restrictions on the types of instances that can be related (i.e., on the class of instances). This is done by defining a *domain* and *range* for an object property.

[42] Similarly to the audio classes introduced in Chapter 4, to refer to tag classes we may use the terms "class" or "category" indistinctly.

[43] In this chapter we use the prefix `fso:` to denote the classes and other definitions of our ontology. The prefix `fso:` stands for "Freesound ontology". However, this is just a convenient name we use to make our explanations more clear.

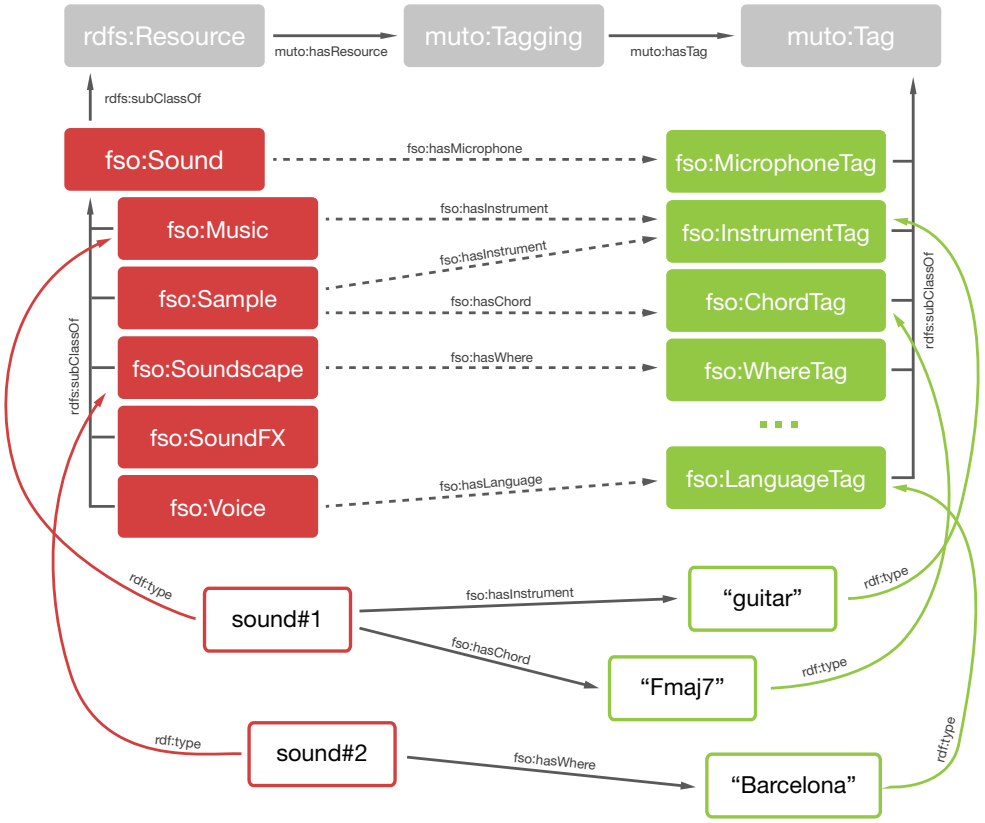


Figure 6.1: Conceptual diagram of the extension of the MUTO ontology that drives the tag recommendation system. Solid boxes represent the classes of the ontology, while arrows indicate object properties. Dashed arrows represent the definitions of the object properties that relate audio categories with tag categories. The boxes at the bottom exemplify tag and resource instances. For the sake of clarity, only a small subset of `muto:Tag` subclasses are displayed in this figure. A complete list can be found in Table 6.1.

categories defined in the ontology is given in Table 6.1. More details on the definition of tag categories and on how we populate them with tag instances are given in Sec. 6.2.2.

Following the same idea of tag categories, we also extend the MUTO ontology by incorporating `rdfs:Resource` subclasses that allow the grouping of resources into a number of audio categories (left side of Fig. 6.1). We define a generic `fso:Sound` subclass for `rdfs:Resource`, and five subclasses for the `fso:Sound` class which correspond to the audio classes that we have already defined in our previous version of the tag recommendation system (Sec 4.2.2). As it can be seen in Fig. 6.1, the five subclasses of `fso:Sound` are named in accordance with their corresponding audio class names.

Finally, we also extend the tagging ontology by defining a number semantic

relations in the form of object properties. The purpose of these object properties is to represent relations between tag and resource instances (dashed lines in Fig. 6.1). Every included object property defines, as its range, a particular `muto:Tag` subclass, and as its domain, at least one of the `rdfs:Resource` subclasses. Therefore, we define as many object properties as tag categories. These object properties are named according to their ranged tag category. For example, the property `fso:hasInstrument` ranges instances of `fso:InstrumentTag`, and its domain includes instances of `fso:Sample` and `fso:Music` audio categories. Therefore, `fso:hasInstrument` can relate sounds belonging to either the `fso:Sample` or `fso:Music` classes, with tag instances of the type `fso:InstrumentTag`. By inspecting the domains and ranges of these object properties, our ontology can determine which tag categories are relevant when describing a resource of a given audio category. For instance, and following the previous example, the ontology can determine that `fso:InstrumentTag` tags, are relevant when annotating sounds of the `fso:Music` or `fso:Sample` audio categories, but that are not so relevant when annotating sounds of other categories such as `fso:Soundscape`. Object properties can be defined with multiple domains, meaning that a particular tag category can be considered relevant to more than one audio category. In Table 6.1 we show, for every tag category, its corresponding object property range, name and domain.

Using this ontology, it is possible to structure the information embedded in a folksonomy, and also to incorporate some meaningful semantic relations that will be used by our tag recommendation system. The ontology is specifically designed to fit the requirements of our use case, but it could be further extended to be capable of representing more classes of resources, tags and other semantic relations. In fact, because the tag categories we define are of type `muto:Tag`, and `muto:Tag` inherits from SKOS[44] class `skos:Concept`, semantic relations between tag instances to represent, for example, synonymy and polysemy, could be easily included (Echarte et al., 2007; Lohmann et al., 2011). An ontology-based tag recommendation system could then take advantage of these relations to refine tag suggestions. Furthermore, semantic relations between resources could also be defined by making resource subclasses inherit from `skos:Concept`. However, the exploration of these possibilities is out of the scope of the work presented in this chapter. Instead, we here show a simple approach in which tag recommendation can be driven by a domain-specific ontology.

6.2.2 Ontology population

So far we have introduced the definition of the ontology that we use to drive our tag recommendation system. We have seen that in using this ontology we are able to determine a number of tag categories that are considered relevant

[44]SKOS: Simple Knowledge Organisation System ontology. `http://www.w3.org/TR/skos-reference`.

Tag category/Object property range	Object property name	Object property domain	Information facet description
fso:ActionTag	fso:hasAction	fso:SoundFX, fso:Soundscape	Physical activities captured in the recording
fso:AgeTag	fso:hasAge	fso:Voice	Age of the speaker or speakers
fso:ArticulationTag	fso:hasArticulation	fso:Music, fso:Sample	Performance or playing technique
fso:ChordTag	fso:hasChord	fso:Music	Music chords present in the recording
fso:DynamicsTag	fso:hasDynamics	fso:Music, fso:Sample	General loudness characteristics
fso:EnvelopeTag	fso:hasEnvelope	fso:Sample	Envelope of a sound at the note level
fso:GearTag	fso:hasGear	fso:Sound	Gear used to generate the sound
fso:GenderTag	fso:hasGender	fso:Voice	Gender of the speaker or speakers
fso:GenreTag	fso:hasGenre	fso:Music, fso:Sample	Music genre
fso:InstrumentTag	fso:hasInstrument	fso:Music, fso:Sample	Instrument names, brands or types
fso:KeyTag	fso:hasKey	fso:Music	Music tonality of the recording
fso:LanguageTag	fso:hasLanguage	fso:Voice	Languages present in the sound
fso:MaterialTag	fso:hasMaterial	fso:SoundFX	Material of sound sources present in the recording
fso:MeterTag	fso:hasMeter	fso:Music	Music time signature information
fso:MicrophoneTag	fso:hasMicrophone	fso:Sound	Microphone names, brands and types
fso:MoodTag	fso:hasMood	fso:Sound	Moods and emotions conveyed by the sound
fso:NoteTag	fso:hasNote	fso:Sample	Music note present in the recording
fso:OnomatopeiaTag	fso:hasOnomatopeia	fso:SoundFX	Phonetic imitations of the sound
fso:ProcessingTag	fso:hasProcessing	fso:Sound	Techniques used to process the recording
fso:RecordingTag	fso:hasRecording	fso:Sound	Recording techniques used to produce the sound
fso:SoftwareTag	fso:hasSoftware	fso:Sound	Software names, brands or types
fso:TempoTag	fso:hasTempo	fso:Music	Tempo information
fso:TypeTag	fso:hasType	fso:Sound	Generic classification of a sound
fso:WhatTag	fso:hasWhat	fso:SoundFX, fso:Soundscape	Sound sources present in the recording
fso:WhenTag	fso:hasWhen	fso:Soundscape	Indication of the moment when the sound was recorded
fso:WhereTag	fso:hasWhere	fso:Soundscape	Indication of the place where the sound was recorded

Table 6.1: Tag categories defined in the proposed ontology and their corresponding object properties.

for a given audio category. Using this information, the recommendation system could guide the annotation process by suggesting potentially relevant information facets to users. However, our tag recommendation system also relies on the ontology for displaying the suggested tags grouped into tag categories. At this point, the defined ontology does not yet contain any knowledge about the categorisation of particular tags into tag categories. Hence, it can not be used to meet the latter requirement. To overcome that limitation, we populate the ontology with a number of tag instances for the different tag categories. In this way, and only for the tag instances that we populate, the ontology is able to tell to which tag category these belong, and the recommendation system can group them accordingly.

The population of the ontology was performed manually, and in parallel with the definition of the different tag categories of the ontology. For the first stage, we selected the 500 most used tags in the folksonomy of Freesound, and built an interface in which we were presented with these tags one by one. The interface allowed us to classify every tag into an existing tag category, or to create a new tag category if no existing categories fitted the tag at hand. We started the process with no predefined tag categories. The consideration of whether a given tag needed a new tag category or not is highly subjective. In general, the goal was to generate broad tag categories that could be easily understood by users in Freesound. We put a special emphasis on tags describing musical properties. Hence, some narrower tag categories were created for this domain (see below). After classifying all tags, we obtained a number of tag categories representative of the 500 most used tags in Freesound, and a number of tags classified under each category. For the second stage, we manually reorganised some of these categories (combining or splitting them into new categories), and also added other categories that we considered were relevant and missing from the resulting list. Then, we were presented again with the 500 most used tags in Freesound, and classified them into the refined set of categories. Because of the ambiguity and unclear meaning of some tags, their classification into tag categories was not a straightforward task. Furthermore, this problem was accentuated because tags were presented one by one and outside of the context of the sounds they were originally assigned to. In some cases, tags were classified to more than one tag category. For example, the tag `piano` can be considered as a tag describing the dynamics of a music recording (`fso:DynamicsTag`), or as a tag describing the name of an instrument (`fso:InstrumentTag`). Tags whose meaning was not clear or did not fit any of the refined tag categories were discarded.

As a result of the whole process, we obtained the set of 26 tag categories shown above in Table 6.1, and 413 of the 500 most used tags in Freesound classified into these categories. For each of the 413 tags, we populated the ontology with a tag instance of the corresponding tag category. Therefore, after the population process, the ontology includes knowledge about the tag category

(or categories) to which each one of these 413 tags belongs. Although 413 tags represents less than 1% of the total number of distinct tags in Freesound, these tags appear in 86% of sound annotations, and are present in 51% of tag applications. Therefore, we can estimate that the populated ontology is able to tell the tag category of roughly one out of two tags introduced by users when annotating sounds.

We mentioned that some of the tag categories we defined are designed with a narrower scope than other categories. In particular, we define a number of very specific tag categories that describe musical properties such as `fso:ChordTag` or `fso:TempoTag`. These categories can not be widely populated following the process described above because, in general, there are only a few tags among the 500 most used tags in Freesound that fit into these categories. This partially happens because there is not a clear agreement in the folksnomy of Freesound on how to annotate this kind of information. For example, in annotating tempo information, it is very common for some users to employ a tag such as `120bpm`, whereas other users indicate the same information with the pair of tags {`120`, `bpm`}, or with a compound tag `120-bpm`. For these particular tag categories, which we refer to as "narrow tag categories", we performed an extra step of population in which we manually produced a list of invented tags (not necessarily chosen from existing tags in the Freesound folksonomy) and added them to the ontology. Hence, for some of the defined tag categories, we created a list of "post-populated tags". Details on the importance of post-populated tags and how are they treated differently in the tag recommendation process are given in Sec. 6.2.3. Besides post-populating narrow tag categories, we applied the same strategy to the `fso:TypeTag` category. As shown in Table 6.1, `fso:TypeTag` tags are intended to classify sounds into general categories. During the population process we classified several tags into this category. Because of their nature, we consider that `fso:TypeTag` tags are particularly important in the annotation of sounds. Hence, for trying to reinforce the agreement in the tags used under this category (see below), we post-populated `fso:TypeTag` with a hand-crafted list of tags. This list includes some of the tags obtained with the normal population process, and some others that were added to create a more complete and coherent list of generic sound type tags. In Table 6.2 we show, for every tag category, the tags that have been post-populated (if any), and up to the 10th most used tags coming from the normal population process.

6.2.3 Ontology-based tag recommendation

The ontology-based tag recommendation system we describe here is built upon the class-based tag recommendation described in Chapter 4. In Fig. 6.2, we show a block diagram of the recommendation system and highlight the components that are added or modified with respect to the class-based recommendation system. We now summarise these additions. In the following sections, we describe in depth the new components of the ontology-based tag recom-

Tag category	Post-populated and normally-populated tags
fso:ActionTag	click, announcement, close, open, walking, drop, squeak, talking, crash, singing
fso:AgeTag	woman, girl, child, baby, children, boy, kids
fso:ArticulationTag	**vibrato**, **tenuto**, **staccato**, **legato**, **extended**, non-vibrato, extended, vibrato
fso:ChordTag	**A**, **C7**, **Fmaj7**, **Am9**, **Em**, **Gm7**, **Asus4**, **FSharp**, **DSharp7**, **FSharpm**, **Bb**, **Eb7**, **G/A**
fso:DynamicsTag	**pianissimo**, **piano**, **mezzo-piano**, **mezzo-forte**, **forte**, **fortissimo**, **midi-velocity-30**, **midi-velocity-64**, **midi-velocity-120**, piano, mezzoforte
fso:EnvelopeTag	**slow-attack**, **fast-attack**, **medium-attack**, **slow-release**, **fast-release**, **medium-release**
fso:GearTag	computer, tape, vinyl, zoom-h2n, zoom, roland, korg, waldorf, virus, atari
fso:GenderTag	male, female, woman, girl, man
fso:GenreTag	ambient, electronic, metal, electro, industrial, techno, house, trance, dance, dubstep
fso:InstrumentTag	drum, synth, electronic, percussion, bass, snare, guitar, kick, acoustic, digital
fso:KeyTag	**A**, **Cmaj**, **Emin**, **FSharpmin**
fso:LanguageTag	english, american, spanish, portuguese, accent
fso:MaterialTag	water, metal, wood, metallic, glass, plastic, paper, air, gas, steel
fso:MeterTag	**4-4**, **3-4**, **6-8**, **9-8**
fso:MicrophoneTag	neumann
fso:MoodTag	horror, scary, funny, suspense, fun, dream, dramatic, tension
fso:NoteTag	**A**, **CSharp**, **Gb**, **A4**, **E3**, **FSharp4**, **Eb5**, **midi-note-35**, **midi-note-40**, **midi-note-64**, **midi-note-80**, **midi-note-127**
fso:OnomatopeiaTag	click, beep, rumble, ring, tone, bang, pop, buzz, bleep, rattle
fso:ProcessingTag	processed, reverb, distortion, echo, remix, unprocessed, filter, synthesis, delay, raw
fso:RecordingTag	stereo, binaural, mono, studio, xy
fso:SoftwareTag	reaktor, vst
fso:TempoTag	**120bpm**, **140bpm**, **60bpm**, 120bpm, 140bpm
fso:TypeTag	**field-recording**, **fx**, **soundscape**, **voice**, **multisample**, **single-note**, **percussive-hit**, **loop**, **music**, **chord**, **chord-progression**, **melody**, **rhythm**, field-recording, loop, 1-shot, drone, soundscape, beat, vocal, glitch, pad, hit
fso:WhatTag	noise, voice, water, birds, machine, wind, people, human, door, ambiance
fso:WhenTag	spring, night, summer, morning, winter
fso:WhereTag	space, nature, industrial, city, street, kitchen, field, station, forest, sea

Table 6.2: Examples of tags populated per tag category. Tags in bold correspond to those tags that were introduced in the post-population step (post-populated tags), while the other tags show up to the 10th most used tags coming from the normal population process (normally-populated tags).

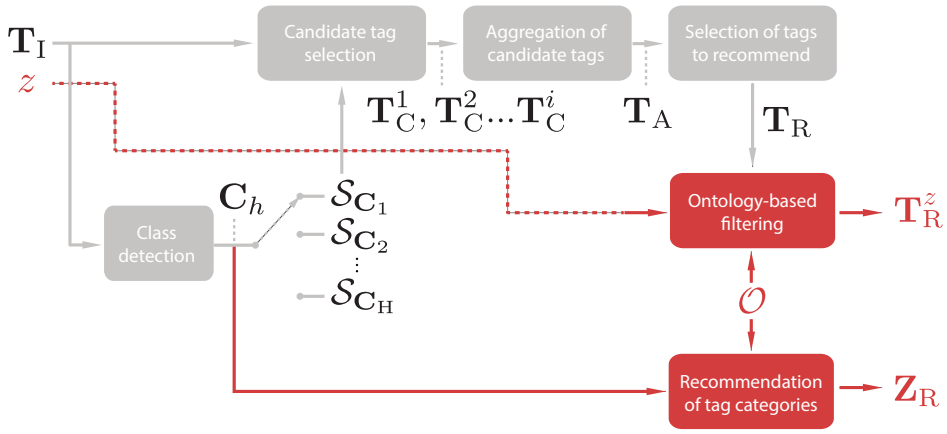


Figure 6.2: Block diagram of the ontology-based tag recommendation system. Note that the ontology-based filtering step takes as input a tag category z, and outputs a set of recommended tags $\mathbf{T}_\mathrm{R}^z$ which depends on z. Also, further note that for generating $\mathbf{T}_\mathrm{R}^z$ and the list of recommended tag categories $\mathbf{Z}_\mathrm{R}$, the recommendation system relies on the ontology that we described in the previous sections (denoted here as $\mathcal{O}$).

mendation system as well as its implementation in terms of user interface.

i) The set of recommended tags depends, as usual, on the input tags $\mathbf{T}_\mathrm{I}$ and on the tag-tag similarity matrix $\mathcal{S}_{\mathbf{C}_h}$ of the audio class $\mathbf{C}_h$ that is selected after the class detection step. However, in the ontology-based system, the recommendation also takes as input a tag category z ($z \in \mathbf{Z}$, where $\mathbf{Z}$ is the set of defined tag categories) chosen by the user. This tag category is used to filter the output of the class-based system and produce the final set of $\mathbf{T}_\mathrm{R}^z$ recommended tags ("Ontology-based filtering" step in Fig. 6.2).

ii) Besides outputting a list of recommended tags, the ontology-based tag recommendation system also produces a set of recommended tag categories $\mathbf{Z}_\mathrm{R}$ that depend on the audio class $\mathbf{C}_h$ that is detected after the audio class detection step ("Recommendation of tag categories" step in Fig. 6.2).

Ontology-based filtering of recommended tags

Given a set of recommended tags $\mathbf{T}_\mathrm{R}$ (as produced by the class-based tag recommendation system) and an input tag category z chosen by the user, the ontology-based system performs the following operation to generate the final set of recommended tags $\mathbf{T}_\mathrm{R}^z$:

$$\mathbf{T}_\mathrm{R}^z = \begin{cases} \mathbf{T}_{\mathrm{Z}'}^z \cup (\mathbf{T}_\mathrm{R} \cap \mathbf{T}_\mathrm{Z}^z) & \text{if } |\mathbf{T}_\mathrm{R}| > 0 \\ \mathbf{T}_{\mathrm{Z}'}^z \cup \mathbf{T}_\mathrm{Z}^z & \text{if } |\mathbf{T}_\mathrm{R}| = 0 \end{cases},$$

where $\mathbf{T}^z_{\mathrm{Z}}$ is the set of normally-populated tags for the tag category z, and $\mathbf{T}^z_{\mathrm{Z}'}$ is the set of post-populated tags for z. As it can be observed, if the first steps of the recommendation system (i.e., Candidate tag selection, Aggregation of candidate tags and Selection of tags to recommend) are able to generate a set of recommended tags $\mathbf{T}_{\mathrm{R}}$, the system filters that set $\mathbf{T}_{\mathrm{R}}$ by discarding all the tags not populated under the tag category z. Then, the post-populated tags for the tag category (if any) are added on to the remaining tags in $\mathbf{T}_{\mathrm{R}}$ (duplicates are removed). On the contrary, if $\mathbf{T}_{\mathrm{R}}$ can not be produced (typically because there are no input tags), the system recommends the union of the post-populated tags (if any) and the normally-populated tags for the corresponding tag category. In this case, normally-populated tags are sorted according to their global frequency of occurrence in the folksonomy of Freesound.

Note that for a given tag category, if there exist post-populated tags in the ontology, these are always recommended in the first place. Therefore, post-populated tags always take a prominent position in the recommendation. With the exception of the tag category `fso:TypeTag`, only narrow tag categories are post-populated (Sec. 6.2.2). The goal of this design choice is that the tags that are post-populated serve more as an example to users that as an actual recommendation. For instance, given the tag category `fso:ChordTag`, the tags which are post-populated provide an idea of how to annotate chords (Table 6.2). Post-population of `fso:ChordTag` includes tags such as `Fmaj7` or `Am9`. These particular chords probably do not suit most of the sounds for which they are recommended, but serve as an example of the syntax that users should follow when introducing chords. Thus, the post-population of tag categories is designed as a way to provide various examples rather than actual recommendations. This concept also applies, to some extent, to the normally-populated tags that are recommended when $\mathbf{T}_{\mathrm{R}}$ is not generated. In that case, tags serve mostly as an example of what kind of information should be introduced in each tag category. However, the post-population of the tag category `fso:TypeTag` is performed to promote the usage of a particular set of predefined tags that categorise sounds into rather broad categories (Table 6.2). By having an extra control over the recommended tags for the `fso:TypeTag` category, we expect that users will annotate the type of their sounds with a more unified vocabulary, using at least one of the tags that we recommend.

Furthermore, note also that the ontology-based tag recommendation system only recommends tags that are populated in the ontology $\mathcal{O}$. Thus, considering that our ontology is populated with a total of 413 unique tags, the recommendation system only recommends tags from that vocabulary of 413 tags. To overcome that limitation, a more comprehensive population of the ontology should be performed (see the discussion in Sec. 6.5).

Recommendation of tag categories

An extra functionality for the ontology-based tag recommendation system is the suggestion of a set of potentially relevant tag categories $\mathbf{Z}_{\mathrm{R}}$ given a set of input tags $\mathbf{T}_{\mathrm{I}}$. This functionality is based on the audio class detection step introduced in the class-based tag recommendation system. Given a set of input tags, the classifier is able to predict to which audio class $\mathbf{C}_h$ a sound belongs to (Sec. 4.2.2). In the ontology-based tag recommendation system, the predicted audio class $\mathbf{C}_h$ is directly mapped to the corresponding audio category defined in the ontology (Sec. 6.2.1). Then, by considering the range of the object properties in $\mathcal{O}$ whose domain matches the predicted audio category, the system is able to generate a list of potentially relevant tag categories $\mathbf{Z}_{\mathrm{R}}$ (see below for an example).

Note that some object properties have as its domain the generic audio class `fso:Sound` (Table 6.1). By inheritance, any instance of the audio categories in the lower level of the hierarchy is also an instance of the class `fso:Sound`. Therefore, these object properties are also considered and their corresponding tag categories are recommended. In fact, tag categories whose object property domain is the audio category `fso:Sound` are always recommended. For example, given a set of input tags whose detected audio category is `fso:Voice`, $\mathbf{Z}_{\mathrm{R}}$ will include all tag categories related with `fso:Sound` (e.g., `fso:MicrophoneTag`, `fso:TypeTag`, `fso:ProcessingTag`, etc.), as well as all tag categories directly related with `fso:Voice` (`fso:LanguageTag`, `fso:AgeTag` and `fso:GenderTag`). Those tag categories more related in first place with the detected audio category (i.e., not by inheritance), are positioned first in the list of recommended categories. We only make an exception with `fso:TypeTag`, which is always placed in the first position to promote its usage.

User interface of the annotation system

The changes introduced in the ontology-based tag recommendation system with respect to the previous class-based system have important implications on the interface for annotating sounds. In Fig. 6.3, we show three screenshots of that user interface. As it can be observed, in an initial stage, the interface provides an input box in which users can start typing tags, and shows the list of all tag categories defined in the ontology (Fig. 6.3a). As soon as users type the first tag in the input box, the tag recommendation system can estimate an audio class $\mathbf{C}_h$, and provide a recommendation of tag categories $\mathbf{Z}_{\mathrm{R}}$. When this happens, the list of tag categories is updated and divided into two parts. The top part lists tag categories in $\mathbf{Z}_{\mathrm{R}}$, and bottom part shows the rest of tag categories (Fig. 6.3b).

Users can introduce tags under the tag categories of the ontology by clicking on the tag category names displayed in the interface. When this happens, the tag category name is appended to the input box, and a pop-over appears

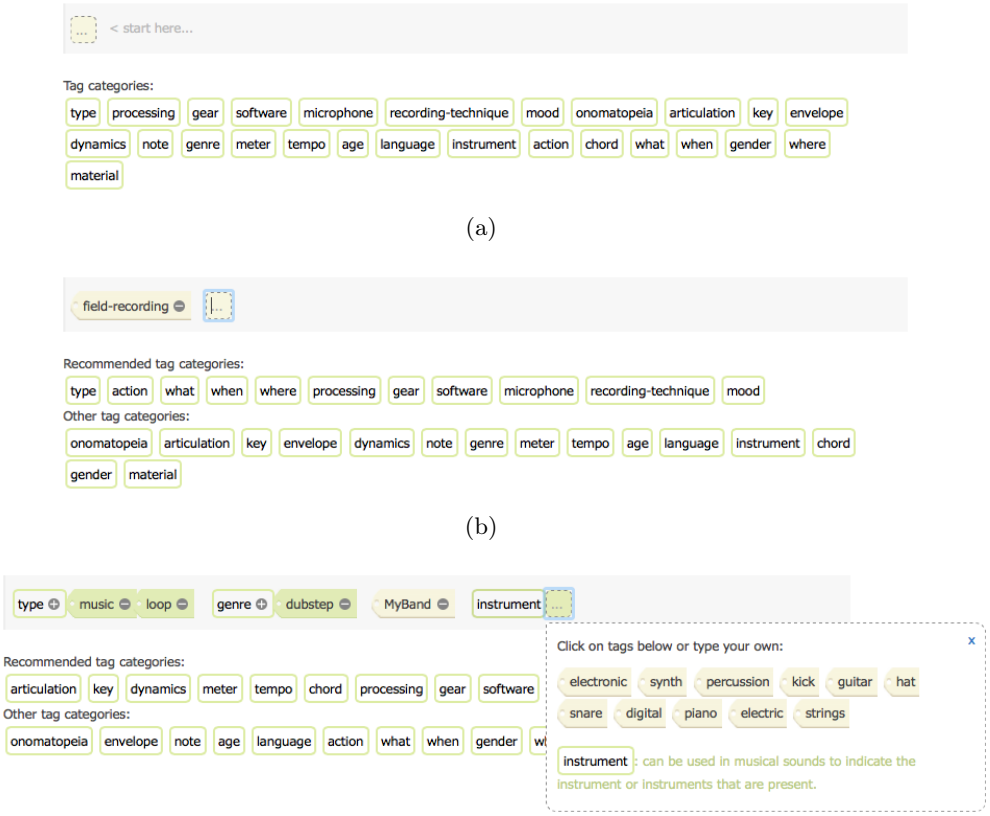


Figure 6.3: Screenshots of the sound annotation interface using the ontology-based tag recommendation system. Screenshot (a) shows the initial state of the interface, with no input tags introduced. Screenshot (b) shows the interface with a single input tag `field-recording`. Screenshot (c) shows the state of the interface with the input tags $\mathbf{T}_{\mathrm{I}}$ ={`music`, `loop`, `dubstep`, `MyBand`}, and showing tag recommendations for the tag category `fso:InstrumentTag`. In the latter example, the user has assigned the tags `music` and `loop` to the tag category `fso:TypeTag`, and the tag `dubstep` to the tag category `fso:GenreTag`. Notice that the tag `MyBand` is introduced with no assigned tag category.

which includes the recommended tags $\mathbf{T}_{\mathrm{R}}^{z}$ (Fig. 6.3c). Similarly to the tag recommendation systems described in previous chapters, recommended tags are sorted according to the scores assigned during the aggregation step of the recommendation process (Sec. 3.2.2), but here we only show a maximum of 20 recommended tags. Users can either choose to add one of the recommended tags or type their own. In either case, the introduced tag is assigned to the selected tag category as an attribute-tag (Sec 6.1). Note that the vocabulary of tags that can be added under a tag category is not restricted. Hence, users can create new tags and assign them to any tag category. Note also that multiple tags can be assigned to a single tag category.

Users are not forced to use any of the recommended tag categories, nor forced to use attribute-tags or click on recommended tags. In fact, users are not presented with any recommended tags until they click on one of the tag category names. In this way, the interface guides the annotation process by suggesting information facets and then providing tag recommendations for every information facet on demand, while at the same time it maintains the flexibility of previous tagging systems and allows users to continue tagging in their preferred way (i.e., without using attribute-tags).

6.3 Evaluation

Here we describe the process we followed to evaluate whether the ontology-based tag recommendation system can better help users to generate comprehensive, coherent, and semantically meaningful resource annotations (when compared to the previous class-based recommendation system of Chapter 4). For that purpose, we designed an online experiment in which participants have to tag a number of sounds from Freesound either using the ontology-based tag recommendation system or the class-based recommendation system. We analyse the logs collected during the experiment and compare both systems by computing a number of metrics. To complement these results, we also perform a second experiment in which the ontology-based tag recommendation system is deployed in Freesound, and collect logs from real-world usage of the recommendation system. These logs are also analysed, when possible, using the same methodology of the previous experiment. In the following sections, we describe the experiments and the analysis metrics we use to evaluate our system.

6.3.1 Description of online experiments

First experiment

The first experiment was carried out during a time period of 22 days from 7 July 2014 to 13 August 2014. The different parts of the experiment are the same as those we used in Chapter 4 for comparing the class-based tag recom-

mendation system with previous systems (see Sec. 4.4). Participants were first presented with a page with the instructions of the experiment and corresponding instructions of the tagging interface. Then, participants had to fill in a questionnaire to collect some basic user data and information about their experience in working with sound libraries, their experience using Freesound and their native language. After completing the questionnaire, participants could start the sound annotation phase. We manually selected a pool of 20 sounds from Freesound, equally distributed in the five audio categories introduced in Chapter 4 (Table 4.1). For each participant in the experiment, we randomly selected 3 sounds per category that had to be tagged. Therefore, participants had to annotate a total of 15 sounds. The tagging interface was assigned at random per participant. Half of the participants used the ontology-based tag recommendation interface, while the other half used the class-based tag recommendation interface. We will refer to the ontology-based tag recommendation interface as ONT, and to the class-based interface as CLA for short. In contrast to the experiment described in Chapter 4, in this experiment sounds were presented to users along with their original textual descriptions from Freesound. This was added to provide more context to participants and facilitate the tagging task (see discussion in Sec. 4.5.5). After annotating the 15 sounds, participants were asked to answer a brief questionnaire to qualitatively evaluate some aspects of the tagging interface. We asked to all participants if the tagging interface was easy to understand, and if the tag recommendations were useful. Furthermore, to participants using the ONT interface, we also asked if tag categories were useful, understandable, and if there was enough variety of tag categories. All questions had to be answered using a 5-point scale, ranging from "strongly disagree" to "strongly agree". Finally, users were given the option to write a comment and provide in this way any other feedback they considered relevant.

Among the 195 participants of the experiment, 109 of them actually completed it. The percentage of participants that completed the experiment is very similar to that obtained in the online experiment described in Chapter 4. On average, we collected 70 alternative taglines for each sound of the aforementioned pool of 20 sounds, half provided using the ONT interface and half provided using the CLA interface.

Second experiment

The second experiment was carried out in Freesound after deploying the ontology-based tag recommendation system as an optional experimental tagging interface. During a period of one week from 18 August 2014 to 25 August 2014, Freesound users were given the option to describe their sounds using the ontology-based tagging interface, labelled as an "experimental tagging interface" (ONT). Otherwise, users could still use the default tagging interface implemented in Freesound which, at the time of the experiment, was the class-

based tagging interface (CLA). The data collected in this experiment came from the usage of both interfaces in a real-world situation. Hence, and as opposed to the first experiment, the sounds to be tagged were those uploaded by the users themselves, and we had no control over them. Because of that, in this experiment, we could not collect multiple taglines per sound. Therefore, some of the analysis metrics that we compute for the first experiment can not be computed for the second experiment.

During the period of the experiment we collected tagging data for 276 sounds and provided by 91 different Freesound users. Almost 70% of the users chose to use the ONT interface. However, not all collected data can actually be included in our analysis. The reason is that the interface of Freesound allows the description of up to 10 sounds at once, meaning that the information we extract from a single annotation session can correspond to the description of multiple sounds. During this process, users may describe sounds in a non-sequential way, and therefore our analysis metrics would not be completely reliable for annotation sessions with multiple sounds. For this reason, we only consider the information coming from these annotation sessions in which only one sound was described. As a result, the data that we finally analyse from the second experiment includes tagging data for 135 sounds, provided by 73 different Freesound users. Such data is evenly distributed between the two interfaces (48% using ONT interface and 52% using CLA interface).

6.3.2 Analysis metrics

To analyse the logs collected in the online experiments and compare ONT and CLA tagging interfaces, we define a number of metrics that we divide in three groups. First, we compute simple quantitative metrics to evaluate aspects such as the time participants spend annotating sounds, the length of the tagline and the number of correctly predicted tags of a given tagline. Then, we perform a more semantic-oriented analysis in which we look at aspects such as the most commonly used tags and tag categories, and define metrics to roughly quantify the comprehensiveness and coherence of annotations. Finally, we analyse the qualitative feedback that participants provide through the questionnaire at the end of the first experiment.

Quantitative metrics

We analyse the collected data using some of the quantitative measures already introduced in Chapter 5. These metrics include the average tagline length (Γ), the average tag application time (Φ_e), and the average number of correctly predicted tags (Ω), formalised in equations 5.7 (Sec. 5.2.3), 5.12 (Sec. 5.2.3), and 4.1 (Sec. 4.4), respectively. In Eq. 5.7, a parameter n is defined to explicitly restrict the set of sounds that are considered for computing the measure on an upload-date basis. In this analysis, we skip this parameter and simply

consider a generic set of resources **R**. Besides these measures, we also include the following two new metrics:

- *Average time per sound*: Similarly to the average tag application time (Eq. 5.12), the average time per sound measures the time required to annotate a sound (i.e., to generate a tagline for a sound), averaged over the sounds described in a set of annotation sessions. In our analysis, every annotation session corresponds to the description of one sound. Hence, average time per sound can be written as

$$\Phi_{\mathrm{r}} = \frac{1}{|\mathbf{A}|} \sum_{a \in \mathbf{A}} \lambda_a, \tag{6.1}$$

  where λ_a is the duration of an annotation session a (in seconds), and **A** is a set of annotation sessions.

- *Average percentage of attribute-tags*: We measure the percentage of attribute-tags found in a tagline (i.e., the number of tags in a tagline that are introduced with a tag category), and average it over a set of resources. Hence, the average percentage of attribute tags can be formalised as

$$\Pi = \frac{100}{|\mathbf{R}|} \sum_{r \in \mathbf{R}} \frac{|\mathbf{T}^r_{\mathrm{T}}|}{|\mathbf{T}^r|}, \tag{6.2}$$

  where $\mathbf{T}^r_{\mathrm{T}}$ is the set of attribute tags of a resource r (i.e., tags with category), $\mathbf{T}^r$ is the set of all tags of a resource r, and **R** is a set of sounds. This metric can be only computed for data collected from the Ont tagging interface, as attribute-tags are a particular feature of that interface.

Semantic analysis

The second part of the analysis we perform is focused on semantic aspects of the annotations. On the one hand, we look at a list of the most common correctly predicted tags for both interfaces (i.e., those tags recommended by the system which are most commonly added to the taglines). In this way, we can have an idea of what kind of tags seem to be more useful as recommendations. On the other hand, we examine which are the most commonly used tag categories in data collected using the Ont interface. From all tags used in every category, we also compute the percentage of them that were actually recommended by the recommendation system (i.e., correctly predicted). In this way, we can have an idea of which are the most useful tag categories and about how effective the recommendation system is in every category. Also, in order to have an indication of whether the tag categories included in the ontology were enough for annotating the sounds, we have a look at those tags that were

introduced without category (non-attribute-tags), and see if these could have been categorised under the existing categories or some of them might require the inclusion of new ones. The latter analyses are also only applicable to the ONT tagging interface.

Besides looking at the previous aspects, we further define two analysis metrics which provide an estimation of the comprehensiveness and coherence of sound annotations generated with both interfaces. These two metrics are based on the analysis of the alternative taglines that we collected for every sound in the first experiment:

- *Annotation comprehensiveness*: Comprehensiveness is measured in terms of the number of distinct information facets that are covered in a tagline. In essence, the more information facets are annotated, the more comprehensive the tagline is considered to be. For each sound annotated in the first experiment, we collected an average of 70 alternative taglines, approximately half of them generated with the ONT interface and half with the CLA interface (Sec. 6.3.1). Considering all the alternative taglines for a given sound, we build an annotation ground truth for that sound, which is then used to evaluate how comprehensive individual taglines are.

  To construct the ground truth for a given sound, we aggregate the individual tags of all alternative taglines into a single set of tags, and manually group them into several information facets. The category of attribute-tags is removed, so that we only add the actual tag to that list. Then, to group the list of tags, we follow a similar process to that described for the definition and population of the tag categories in the ontology (Sec. 6.2.2). Given that, in this case, the grouping of tags is performed with the actual sound as a reference, tags can be grouped according to the kind of information they describe in the context of that particular sound. Hence, the process becomes simpler and less ambiguous. As a result, we obtain a number of information facets with a set of tags classified into them. Given the ground truth and a tagline for a sound, the comprehensiveness of a tagline is computed by comparing the number of information facets covered by the tagline (i.e., the number of facets from the ground truth for which there is at least one tag in the tagline), with the total number of possible information facets defined in the ground truth. Hence, given the intermediate function

$$covered(\mathbf{X}, \mathbf{Y}) = \begin{cases} 1 & \text{if } |\mathbf{X} \cap \mathbf{Y}| > 0 \\ 0 & \text{if } |\mathbf{X} \cap \mathbf{Y}| = 0 \end{cases},$$

  annotation comprehensiveness is defined as

$$\Lambda = \frac{1}{|\mathbf{G}^r|} \sum_{\mathbf{T}_{\mathrm{IF}} \in \mathbf{G}^r} covered(\mathbf{T}^r, \mathbf{T}_{\mathrm{IF}}), \tag{6.3}$$

where $\mathbf{T}^r$ is the tagline of a resource r, $\mathbf{G}^r$ is the ground truth of a resource r, and $\mathbf{T}_{\text{IF}}$ is a set of tags corresponding to one of the information facets defined in $\mathbf{G}^r$. Using this measure, we can estimate how comprehensive are, individually, each of the alternative taglines provided for every annotated sound. Then, we can compare the annotation comprehensiveness of ONT and CLA interfaces by averaging this measure over all the taglines generated with each interface.

- *(In)coherence in annotations*: To evaluate the coherence of a set of alternative taglines for a given sound we, in fact, define a measure of incoherence. Incoherence is measured in terms of the variety of tags that are used to convey a unique semantic meaning (or audio property). The more variety of tags is used to describe a particular property in a set of taglines for a given sound, the more incoherent these taglines are considered to be. For example, given an audio recording of a musical instrument playing a chord, a set of taglines that feature tags like `DMaj`, `D-major`, `Dmajor` and `DM` to indicate the chord, is more incoherent than another set of taglines in which the chord is always indicated with either `DMaj` or `Dmajor` (using fewer variations). In a sense, the incoherence in annotations reflects the agreement on how to annotate particular audio properties among the participants that generated the taglines. Similarly to annotation comprehensiveness, we estimate the incoherence in annotations by building an annotation ground truth for each of the sounds annotated in the first experiment. Then, this ground truth is used to estimate the incoherence of the taglines of the corresponding sound. Hence, incoherence in annotations can only be computed on data collected in the first experiment.

  The annotation ground truth that we build for this measure is different to that built for annotation comprehensiveness. Here, given a sound, we also aggregate the individual tags of all alternative taglines into a single set of tags, but instead of grouping them into rather broad information facets, we group together those tags that convey the same or very similar information. For example, we group together tags like {`Funk`, `funk`, `funky`}, which denote the same music genre, tags like {`distortion`, `smooth-distortion`, `distorted`,`overdrive`}, which all describe a very similar audio processing effect, or tags like {`talking`, `speak`, `talk`, `chatting`, `speaking`}, which refer to the same activity. Tags with no other equivalent (or very similar) tags are not considered in the ground truth. As a result, the annotation ground truth for a sound consists of a list of sets of tags that are almost equivalent from a semantic point of view (i.e., a list of *synsets*[45]). Considering the ground truth for a sound and a set

[45]To simplify explanations, we adopt the terminology used in WordNet (Miller, 1995), which refers to a set of synonyms as a "synset". In our case, we use the term synset with a slightly broader meaning, as tags are not strictly grouped because of being synonyms, but because of featuring high semantic equivalence.

of alternative taglines, we iterate over its identified synsets and count how many of the tags in the synset are present in the set of taglines. Averaging that value over the different synsets, we obtain an indication of the incoherence in a set of taglines. The measure can be formalised as

$$I = \frac{1}{|\mathbf{G}^r|} \sum_{\mathbf{T}_{\mathrm{SY}} \in \mathbf{G}^r} |\mathbf{T}^r_{\mathrm{TL}} \cap \mathbf{T}_{\mathrm{SY}}|, \tag{6.4}$$

where $\mathbf{T}^r_{\mathrm{TL}}$ is the union of all tags from a set of taglines of a resource r, $\mathbf{G}^r$ is the annotation ground truth of a resource r, and $\mathbf{T}_{\mathrm{SY}}$ is a set of tags corresponding to one of the synsets defined in $\mathbf{G}^r$. Using this measure, we can estimate how incoherent a set of taglines is for a given sound. By computing this measure over sets of taglines collected with the Ont and Cla interfaces, we can compare them in terms of incoherence in resulting annotations.

Qualitative feedback

To analyse the qualitative feeback provided by users we compute the average over the responses of the different questions answered in the 5-point scale (see Sec. 6.4.3), and relate these with the results of the quantitative and semantic analysis. Furthermore, we compare the responses of these questions which are common to participants using both interfaces. Finally, we comment on the extra qualitative feeback that we collected in the form of textual comments.

6.3.3 Analysis methodology

The analysis we perform is mainly centred on the data collected from the first experiment. This is because, as we have seen, the nature of the second experiment does not allow us to compute all the metrics listed above. Hence, the analysis of the data from the second experiment is used, whenever available, as a complement to the analysis of the first experiment. Nevertheless, with the exception of some metrics, we use the same methodology to analyse the data collected from both experiments. In general, we consider data generated with the Ont interface separated from data generated with the Cla interface, and then compare the results of the different metrics. Statistical significance is assessed using the Mann-Whitney U test with a significance level of 0.05 (Corder & Foreman, 2009).

Before computing the analysis measures described above, we filter the collected data to remove potentially irrelevant or noisy logs. The filter is applied at the sound level, meaning that we discard the information from individual sounds that do not pass the filter. The first filter we compute operates on the duration of annotation sessions for sounds. We discard annotation sessions based on the interquartile range of their duration. Let q_1 be the first quartile and q_3 be the third quartile of the durations of annotation sessions, we discard sounds whose

annotation session duration is outside the range $[q_1 - 3(q_3 - q_1), q_3 + 3(q_3 - q_1)]$. In practice, this means that we only consider data corresponding to sounds that were annotated in less than 260 seconds. We apply a second filter in which we discard all sound annotation sessions in which there are no logged *calls* to the tag recommendation system. In these sessions, no tag recommendations were provided to the user. This might happen because of different reasons (see below).

After applying the filter to all collected annotation sessions of the first experiment, we see that 69% pass the filter. Most of the annotation sessions comply with the timing restriction (97%), but only 72% feature at least one call to the recommendation system. Annotation sessions that do not pass the filter mostly correspond to sessions using the ONT interface. This observation can be partially explained because, when using the ONT interface, tags are only recommended when users click on the tag categories (Sec. 6.2.3). Conversely, when using the CLA interface, tags are automatically recommended when users start typing. As a consequence, it can happen, particularly when using the ONT interface, that no tags are recommended at all. We observe that only 42% of annotation sessions using the ONT interface feature at least one call to the recommendation system. Possible reasons for that could be that users do not properly understand how the interface works or that they do not think that tag categories are meaningful or necessary. Overall, the fact that less than half of the annotation sessions using the ONT interface pass both filters indicates that a lot of participants, for whatever reason, did not take advantage of the features of the interface as we expected. When applying the filter to the data from the second experiment, we observe practically the same percentages. Therefore, one interesting aspect of the ONT interface that should be further investigated is why such a big percentage of users does not take advantage of recommendation functionalities as we expected (see discussion in Sec. 6.5).

6.4 Results

6.4.1 Quantitative metrics

Average time per sound

The average time per sound Φ_r reveals that sound annotations tend to require more time when using the ONT interface. In particular, we observe a statistically significant average increase of 48 seconds ($p = 5.28 \cdot 10^{-42}$). Fig. 6.4 shows the probability density function of the time per sound for both interfaces. The average time per sound is of 60 seconds for participants using the CLA interface, and of 109 seconds for participants using the ONT interface. Similarly, when looking at real-world data from the second experiment, we also observe that users that choose the ONT interface spend more time annotating their

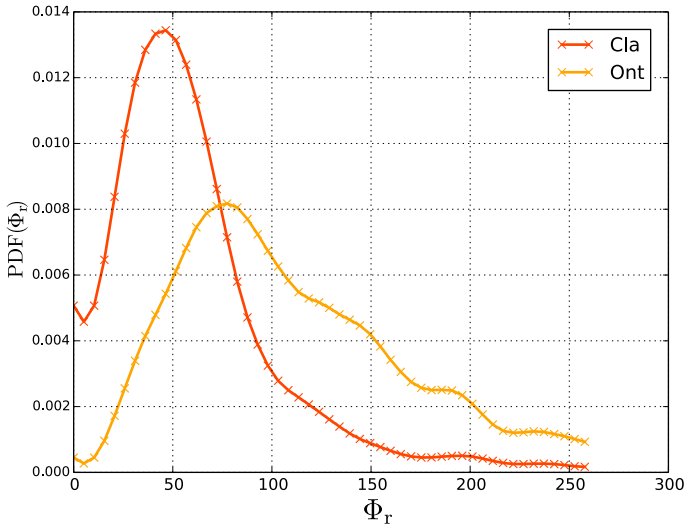


Figure 6.4: Probability density function of the time per sound Φ_r with ONT and CLA interfaces. Curves are smoothed using a Hann window of 11 points.

sounds. In this case, the average time per sound of both interfaces is significantly higher than in the first experiment (156 seconds for CLA and 301 seconds for ONT), and so it is the difference between them (145 seconds, $p = 8.27 \cdot 10^{-5}$). This can be explained because, in the second experiment, users do not only have to annotate sounds, but also have to provide other information which is required by the Freesound uploading interface. This information includes a name for the sound, a textual description, a license and, optionally, geolocation and pack[46] information. Furthermore, the increased difference between both interfaces observed in the second experiment can also be due to users experimenting with the interface to understand how it works. In the first experiment, before starting to annotate sounds, users were provided with instructions for the corresponding interface. Conversely, in Freesound, these instructions were not provided beforehand, and could only be accessed through a link that was provided during the annotation process.

Average tag application time

The average tag application time Φ_e features a very similar behaviour to that observed with Φ_r. Participants of the first experiment need, on average, 11 seconds for assigning a tag using the CLA interface, and 20 seconds using the ONT interface. The increase of 9 seconds is statistically significant ($p = 2.82 \cdot 10^{-41}$). Similarly, in the second experiment, both the increase of the average tag application time is higher (28 seconds, $p = 1.34 \cdot 10^{-6}$), and also the average times for both interfaces (19 seconds for CLA and 47 seconds for

[46] In Freesound, sounds can be explicitly grouped in packs that users define.

ONT). Besides being in line with the previous measure, these results suggest that the extra time that users require for annotating sounds with the ONT interface is not because the generated taglines are longer, but because the assignment of each tag requires more time.

Average tagline length

The observations on the average tagline length that we can make by analysing the data from the first experiment confirm the previous suggestion that taglines do not tend to be longer for participants using the ONT interface. Although the average tagline length is slightly higher (6.61 tags for ONT interface and 6.24 tags for CLA interface), the difference is not statistically significant. However, the results of the average tagline length computed for the second experiment are surprising. In this case, we observe that taglines using the CLA interface are in fact longer that those generated with the ONT interface (12.07 tags and 9.10 tags respectively, $p = 2.45 \cdot 10^{-2}$). The reason for this behaviour is not clear. One possible explanation is that with the usage of attribute-tags, sound annotations may appear to be more specific and complete while using fewer tags. Hence, users might feel that the description is good enough using fewer tags. In practice, the ONT interface merges tag categories and actual tags in the same space (see Fig. 6.3), and this might cause the perception that the tagline is longer than it actually is.

Average number of correctly predicted tags

The analysis of the average number of correctly predicted tags does not reveal a significant difference when comparing interfaces. Sounds annotated with the CLA interface feature an average of 2.48 correctly predicted tags, which is very similar to that obtained in our previous evaluation of the CLA recommendation method (Chapter 4, Table 4.4), and of the impact of the tag recommendation system (Chapter 5, Sec. 5.3.3). Sounds annotated with the ONT interface feature a slightly higher average of 2.60 correctly predicted tags, but the difference is not statistically significant ($p = 1.22 \cdot 10^{-1}$). On the data collected for the second experiment, we again observe similar results with no statistically significant differences between interfaces (average of 3.30 and 2.97 correctly predicted tags for ONT and CLA interfaces, respectively, $p = 3.56 \cdot 10^{-1}$). What these results suggest is that the grouping of tags into tag categories and the recommendation of post-populated tags provided by the ONT interface does substantially not impact the number of tags from taglines that are correctly predicted by the recommendation system.

Similarly to what we did in Chapter 4, we analysed if there is a correlation between the average number of correctly predicted tags and users' expertise and language (Sec. 4.5.1). In particular, we consider the results from the first experiment and divide collected logs between groups of experienced and

non-experienced participants, and native and non-native English speakers. Accordingly to what we describe in Sec. 4.5.1, we observe here small increases in the number of correctly predicted tags for both expert and native speaker participants (in both interfaces). However, as opposed to the previous analyses, the differences in this case do not appear to be statistically significant. Further research should be carried out in order to make stronger claims regarding that matter.

Average percentage of attribute-tags

Considering the taglines generated with the ONT interface for our first experiment, we observe that an average of 72% of the tags are introduced under a tag category (i.e., are attribute-tags). In the second experiment, this percentage is slightly lower, at 65%. This means that a significant number of tags from taglines are contextualised in one of the defined tag categories of the ontology, and therefore their semantic value is effectively improved. However, it is important to note that these percentages are achieved when considering filtered data as described in Sec. 6.3.3. This filter removes, among others, data from annotation sessions in which participants do not take advantage of the tagging interface as we expected. In Fig. 6.5, we show the histogram of the number of attribute tags per tagline when considering unfiltered data for the first experiment. What we observe now is that more than half of the taglines (59%) feature no attribute-tags, whereas other taglines tend to have between 1 and 7 attribute-tags. A similar observation can be made in the second experiment, with 43% of taglines featuring no attribute-tags. These results suggest that the tagging interface should better encourage the usage of attribute-tags.

To complement these results, we analyse if there is a correlation between users' expertise and the percentage of attribute-tags. We observe that experienced users tend to include, on average, 13% more attribute-tags than non-experienced users, that difference being statistically significant ($p = 3.86 \cdot 10^{-2}$). This might be explained because experienced users better understand the advantages that attribute-tags provide in terms of description accurateness and further retrieval possibilities.

6.4.2 Semantic analysis

Most common correctly predicted tags

The most common correctly predicted tags for both experiments and interfaces are listed in Table 6.3. Looking at the tags of the first experiment, we observe that, regardless of the interface, an important number of tags have a semantic meaning that would belong to the `fso:TypeTag` category (e.g., `field-recording`, `voice`, `ambiance`). Other common tags belong to well defined musical concepts (e.g., `loop`, `synth`, `electronic`, particularly in the ONT interface). Overall,

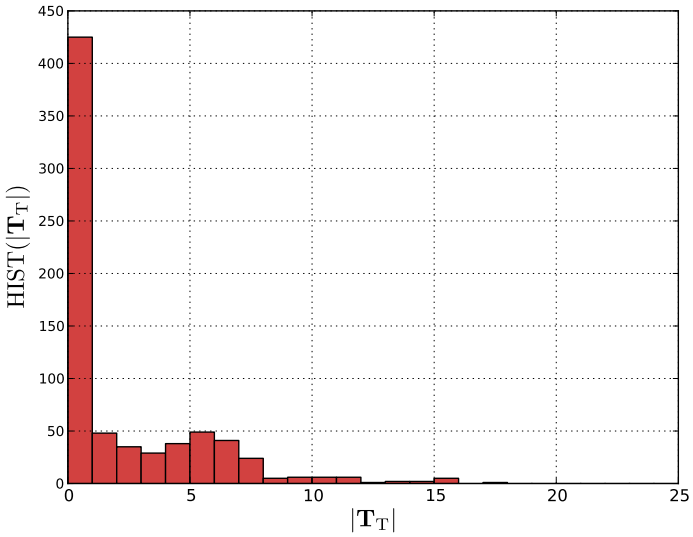


Figure 6.5: Histogram of the number of attribute-tags per tagline $|\mathbf{T}_T|$ in the first experiment (including unfiltered data).

Experiment	ONT **interface**	CLA **interface**
First	field-recording, voice, synth, condenser, single-note, fx, loop, soundscape, stereo, ambiance, 120bpm, electronic, percussive-hit, city, mezzo-forte, fast-attack, processed, mono, male, rhythm	field-recording, electronic, loop, synth, voice, people, rain, male, bell, child, weather, female, ring, nature, bells, vocal, beat, ambience, writing, city
Second	fx, soundscape, field-recording, talking, horror, condenser, stereo, male, english, voice, loop, summer, atmosphere, neumann, crowd, pop, distortion, suspense, female, violin	drums, drum, atmosphere, fx, ten, dance, house, ambiance, gloomy, deep, down, cavern, ambient, rave, chime, airplane, electronic, monster, sub, 0

Table 6.3: First 20 most common correctly predicted tags for the first and second experiments and for ONT and CLA interfaces.

both interfaces present similar (or very related) most common correctly predicted tags. In fact, 20% of the first 50 most common correctly predicted tags in ONT and Cla interfaces are *exactly* the same. This can be explained because the pool of sounds to annotate in the first experiment was limited to 20 sounds, and the concepts to annotate are determined by the sounds. However, of particular interest is the case of very specific tags such as `120bpm`, `mezzo-forte` and `fast-attack`, which are included in the list of most common correctly predicted tags for the ONT interface. These tags are post-populated under the `fso:TempoTag`, `fso:DynamicsTag` and `fso:EnvelopeTag` categories respectively (Table 6.2.2), and are therefore always recommended when using

the ONT interface. If these tags were not recommended, users would presumably employ different variations of the same tags (e.g., `mezzoforte` instead of `mezzo-forte`) that would prevent them from being amongst the most common correctly predicted tags unless these were very obvious. In fact, we hypothesise that this is what happens for the tags that belong to the `fso:TypeTag` category which, as we commended before, are an important part of the most common correctly predicted tags in both interfaces. We can see, for example, that the tag `voice` takes the second position in the ONT interface, and that the other tags in the list have completely different meanings. Interestingly, in the CLA interface, we see how two tags that present a notable semantic overlap (`voice` and `vocal`), are both in the list of most common correctly predicted tags (in the fifth and sixteenth position).

If we look at the most common correctly predicted tags of the second experiment, we observe some differences. Here, we do not observe such similarity between tags from both interfaces (only 8% are common amongst the first 50 most common correctly predicted tags). This was to be expected, as the sounds described in this experiment are not controlled, hence the potentially relevant concepts to annotate are not necessarily comparable between interfaces. However, in the tags from the ONT interface, we still observe a great presence of post-populated tags from the `fso:TypeTag` category, which is not observed in the CLA interface (e.g., `voice`, `soundscape`, `field-recording`). Overall, these results suggest that `fso:TypeTags` tags are more useful as tag recommendations in the ONT interface, and that post-population in general seems to contribute in the coherence of the vocabulary.

Usage of tag categories

In this section we have a look at the most commonly used tag categories in both experiments, and at the percentage of the tags introduced in every category that were correctly predicted by the recommendation system (Table 6.4). Usage is computed as the percentage over the total number of taglines (of the ONT interface) that feature at least one attribute-tag of a particular tag category. We observe that the most used tag categories are, in general, those which are applicable to virtually any kind of sound (first rows of Table 6.4). The tag categories that are highly used depend on the kinds of sounds being annotated (e.g., music sounds require music-related tag categories). Therefore, the comparison between tag categories usage for both experiments is not a priori very meaningful. However, we observe that there is a significant correlation between the ranking of tag categories usage in both experiments (second and third columns of Table 6.4). To asses this correlation, we employ the Spearman's rank correlation coefficient (Corder & Foreman, 2009) and observe a correlation coefficient of $\varrho = 0.745$ with a p-value of $p = 1.24 \cdot 10^{-5}$. We hypothesise that this correlation can be explained by the presence of generic tag categories that can be relevant to different kinds of sounds. Overall, it

is interesting that in both experiments, the `fso:TypeTag` is the most used tag category. As we explained before, our design puts a special emphasis on the `fso:TypeTag` category (Sec. 6.2.2). Hence, the broad presence of this category in sound descriptions is one successful outcome of using the ONT interface.

Another aspect that we examine is the percentage of correctly predicted tags introduced under every tag category. To evaluate this, for every tagline we take into account all introduced tags from each tag category and compute the percentage of these that were recommended by the system during the annotation session. This value is then averaged over all taglines (fourth and fifth columns of Table 6.4). A high percentage indicates that many of the tags used under a category come from system recommendations. Again, we observe that there is a high correlation between the percentage of correctly predicted tags per category in both experiments ($\varrho = 0.906$, $p = 2.19 \cdot 10^{-7}$), meaning that tag categories feature similar percentages in both experiments. Particularly relevant are those categories in which both the percentage of usage and the percentage of correctly predicted tags are high (Table 6.4). In these cases, it can be hypothesized that the ONT interface successfully contributes in the homogenisation of the information facet of the particular tag category, as users reuse tags suggested by the recommendation system rather than creating new ones. This is the case of the tag categories in the first rows of Table 6.4, and particularly of the `fso:TypeTag` category, which shows a wide usage and wide reuse of the tags recommended by the system.

Most commonly used tags without tag category

In Table 6.5 we list the most commonly used tags in the taglines generated with the ONT interface that are introduced without any tag category (non-attribute-tags). By examining these tags, we expected to observe some patterns of tags without category that could suggest the need of adding new categories. However, what we observe is that in both experiments, most of the tags without category could be easily categorised into one of the tag categories defined in the ontology. Hence, there does not seem to be a particular reason (related with available tag categories) about why these tags were not introduced as attribute-tags. Possible explanations are that users simply do not feel the need or see the advantages of using tag categories, or that the meaning of tag categories is not clear enough so that users can easily introduce tags under them. Hence, although the ontology could be more comprehensive and include more tag categories, our early results do not suggest that the current number of categories is a limitation for the ONT interface.

Annotation comprehensiveness

As described above, we measure how comprehensive sound annotations are by estimating the number of information facets that are covered in a tagline of

	% usage		% correctly predicted	
Tag category	**First exp.**	**Second exp.**	**First exp**	**Second exp.**
fso:TypeTag	61.40	50.00	93.36	100.00
fso:InstrumentTag	36.76	21.88	58.63	80.00
fso:WhatTag	22.79	28.12	53.67	69.44
fso:MicrophoneTag	20.96	28.12	40.35	66.67
fso:RecordingTag	18.01	40.62	85.14	85.71
fso:ProcessingTag	18.01	12.50	65.40	100.00
fso:MoodTag	14.71	18.75	42.86	40.00
fso:GearTag	13.97	18.75	17.54	33.33
fso:ActionTag	13.60	25.00	51.35	85.71
fso:WhereTag	12.87	21.88	50.65	44.44
fso:GenderTag	11.40	18.75	90.38	100.00
fso:NoteTag	11.03	0.00	28.33	-
fso:TempoTag	10.66	3.12	44.83	-
fso:SoftwareTag	9.56	21.88	1.96	25.00
fso:OnomatopeiaTag	9.19	0.00	41.67	-
fso:MaterialTag	8.82	9.38	77.50	100.00
fso:DynamicsTag	7.72	0.00	100.00	-
fso:EnvelopeTag	7.72	3.12	100.00	100.00
fso:AgeTag	6.99	6.25	66.67	-
fso:LanguageTag	6.62	15.62	50.00	75.00
fso:KeyTag	5.51	0.00	13.33	-
fso:GenreTag	4.78	9.38	33.33	50.00
fso:ArticulationTag	4.41	0.00	83.33	-
fso:WhenTag	3.31	12.50	41.67	50.00
fso:MeterTag	2.94	3.12	100.00	100.00
fso:ChordTag	2.57	0.00	42.86	-

Table 6.4: Percentage of usage of the different tag categories in the first and second experiments, and percentage of correctly predicted tags for every category and experiment. Tag categories are sorted according to their percentage of usage in the first experiment. Note that this table only includes data gathered with the Ont interface, as the concept of tag categories is not present in the Cla interface.

First experiment	Second experiment
loop, bass, bells, synth, voice, dark, cackle, field-recording, metal, piano, restaurant, bell, child, synthesizer, ambience, sample, percussion, talking, note, radio	beatboxing, percussion, beats, vocalpercussion, beatbox, drums, beat, SFX, vocal, male, erra, draw, detroit, dj, dark, ding, cylinder, Soundeffects, shake, scratch

Table 6.5: 20th most commonly used tags without tag category (non-attribute-tags) for the first and second experiments. Note that this table only includes data gathered with the ONT interface, as the concept of tag categories is not present in the CLA interface.

a given sound, compared to the total number of potentially relevant information facets for that sound (Sec. 6.3.2, Eq. 6.3). Our results show that taglines generated with the ONT interface cover, on average, 23% of the potentially relevant information facets, while taglines generated with the CLA interface cover an average of 18% of the facets. Hence, we observe a statistically significant average increase of 5% when using the ONT interface ($p = 1.11 \cdot 10^{-3}$). This increase suggests that, by recommending tag categories to users, the ONT interface effectively helps in generating more comprehensive sound annotations that cover more information facets. However, even if that increase is statistically significant, we have to take into account that it is computed by considering the filtered set of annotation sessions. Hence, the goal of improving annotation comprehensiveness is only partially achieved in some annotation sessions.

Incoherence in annotations

On average, taglines generated with the ONT interface report an incoherence value of $I = 2.05$, while taglines generated with the CLA interface show an incoherence of $I = 2.95$. The difference is statistically significant, with $p = 3.73 \cdot 10^{-8}$. These results suggest that, considering all alternative taglines for a given sound, those generated with the ONT interface feature an average of 2 distinct tags to refer to semantically similar concepts (e.g., among the taglines genenerated with the ONT interface, we see that two tags like `Thunderstorm` and `thunder-storm` are used to refer to the concept of a "thunderstorm"), while taglines generated with the CLA interface feature an average of 3 distinct tags (e.g., following the previous example, in taglines generated with the CLA interface we might find a third variation for the "thunderstorm" concept such as `rainstorm`). Hence, taglines generated with the CLA interface tend to be less coherent than taglines generated with the ONT interface, as the way in which concepts are tagged is less unified across sounds.

A possible explanation for the observed difference in I is the contribution of post-populated tag categories in the ontology. The tags recommended for these categories act more as example-tags that suggest to users how to de-

scribe particular information facets like those represented by `fso:NoteTag` or `fso:DynamicsTag` (see Sec. 6.2.2). In these categories, it is likely that users would employ different tag variants to describe a single concept. For example, users would probably employ different naming conventions to indicate a musical note. However, by being exposed to the example tags provided by the ontology-based interface, a particular naming convention is suggested and alternative variants are potentially reduced. This seems to be particularly true for the case of tags introduced under the `fso:TypeTag` category. The `fso:TypeTag` category is widely used in the sound annotations collected in our experiments, and features a high percentage of correctly predicted tags (Table 6.4). This ensures that most of the sounds are given at least one known tag describing their "type". Thus, the "type" property is annotated coherently across sounds. Notice however that the interface allows users to introduce new tags (i.e., not recommended) under any of the tag categories. Hence, the particular case of the `fso:TypeTag` category is an example that the ontology-based tagging interface can achieve a successful level of tagging coherence across sounds without limiting the flexibility of creating new tags.

6.4.3 Qualitative feedback

In this section we report the qualitative feedback that we gathered through the questionnaire at the end of the first experiment. In Table 6.6 we show the average answers to the questions that were asked. Questions had to be answered on a standard 5-point scale. We normalised the responses so that a value of 1 corresponds to "strongly agree" and a value of 0 corresponds to "strongly disagree". We report the average answers for the set of all participants that finished the experiment ("Not filt." column), and for the set of participants whose annotated sounds comply with the filter described in Sec. 6.3.3 ("Filt." column). In general, we observe that, qualitatively, both interfaces are considered to be rather easy to understand, with no statistically significant differences. However, participants using the CLA interface consider that tag recommendations were more useful than participants using the ONT interface, with an statistically significant increase between 0.09 ($p = 2.93 \cdot 10^{-2}$) and 0.15 ($p = 7.06 \cdot 10^{-4}$) for the non-filtered and filtered set of participants, respectively. Furthermore, average responses for the questions regarding usefulness, variety and understandability of tag categories report lower scores, generally in the range corresponding to "Neither agree nor disagree" and "Agree". Interestingly, we can see that these scores are a bit higher when only considering the filtered set of participants. This can be explained because this set only includes participants who, a priory, took advantage of the functionalities of the ONT interface as we expected (Sec. 6.3.3).

Besides the previous questions, participants in the first experiment were also given the option to provide further feedback in the form of textual comments. In general, comments were positive about both interfaces. Some participants

Question	Ont interface		Cla interface	
	Not filt.	Filt.	Not filt.	Filt.
The tagging interface was easy to use	0.75	0.78	0.78	0.78
Tag recommendations were helpful during the annotation process	0.66	0.72	0.81	0.81
Tag categories were useful as a guide for the annotation process	0.64	0.72	-	-
The number and variety of tag categories was enough to annotate the sounds	0.62	0.66	-	-
The meaning of tag categories and the type of tags that should be used in every category was easy to understand	0.62	0.68	-	-

Table 6.6: Questionnaire responses. Response values are normalised so that a value of 1 corresponds to "strongly agree", and a value of 0 corresponds to "strongly disagree".

included suggestions of new features to improve the interfaces such as auto-completion of tags and displaying more tag recommendations. Interestingly, one participant that completed the experiment using the Cla interface, suggested that recommending a predefined taxonomy of audio categories could help in the annotation process (similarly to what the tag category `fso:TypeTag` does in the Ont interface). Other users commented that tag recommendations in the Ont interface were too few, probably not understanding that the tags to be introduced under every category were not limited to the recommended ones.

6.5 Conclusion and discussion

In this chapter we have explored a new perspective on tag recommendation systems which combines the use of a folksonomy and a domain-specific ontology to guide the annotation process and recommend tags. By combining the use of a folksonomy and an ontology, the resulting annotation system is expected to gather better structured resource annotations, while being able to maintain the flexibility and ease of use of standard tagging systems. We described the design of an ontology tailored to the needs of our recommendation system, and explained how the system takes advantage of that ontology to recommend tags and tag categories depending on a set of input tags. Using a tag recommendation system such as the one described in this chapter, we expect users to provide more coherent, comprehensive and semantically meaningful sound annotations. The system we propose has been evaluated with two online experiments, one in a controlled environment and another one in the real-world context of Freesound. In addition, it has been compared with the class-based tag recommendation system that we described and evaluated

in previous chapters.

The analysis performed in both experiments yields similar results, and shows that the ontology-based interface can, in some cases, contribute to the improvement of sound annotations. In particular, we distinguish between two usage patterns. On the one hand, we observe that users who spend enough time for annotating sounds and take advantage of the functionalities of the ontology-based interface, are able to provide better sound annotations. These sound annotations tend to cover more information facets (i.e., are more comprehensive), tend to use less variants of tagging concepts (i.e., are more coherent), and their tags are more semantically meaningful (i.e., tags are introduced under tag categories). However, on the other hand, we observe that approximately 55% of users do not actually take advantage of the functionalities of the ontology-based interface. Users annotating sounds with the ontology-based interface that do not click on any of the tag categories are not recommended with any tags at all. Therefore, in these cases, the benefits of tag recommendation are lost and the interface might even have a negative impact on the resulting sound annotations as compared to annotations performed with the class-based tag recommendation system.

One possible explanation as to why the majority of users did not use the ontology-based tagging interface as we expected is that the interface itself was hard to use and understand. However, according to the qualitative feedback provided by participants of the first experiment, this is probably not the main reason. Another possible explanation is that the concept of tag categories and the particular categories defined in the ontology were not meaningful to users. Again, we believe this is not the main limitation of the tagging system as we have shown that the kinds of tags introduced without categories could have been introduced under existing tag categories, and, according to the qualitative feedback, users moderately agreed in that the set of categories was sufficient. For these reasons, we believe that the main explanation for the timid usage of the ontology-based tagging system is that the interface does not promote enough the use of tag categories and, in general, the benefits of accurate sound descriptions for further retrieval and reuse. Hence, further research could be aimed at understanding what kind of mechanisms could be used to better promote that aspect and facilitate the usage of the interface. For example, a minimum number of attribute-tags could be set as mandatory, or users could be given some sort of reward when generating longer descriptions including more attribute-tags (i.e., users could be given a score that would be public to other users). Furthermore, and in the particular case of sound sharing, content-based strategies could be used to automatically predict the tags that could be added under some of the narrow tag categories such as `fso:NoteTag` or `fso:TempoTag`, and pre-fill the annotation with these predictions.

Another aspect of the ontology-based tag recommendation system that can not be compared favourably to the class-based system is the usefulness of

tag recommendations. We observed that there is no statistically significant difference in the number of correctly predicted tags when comparing both interfaces, and that users found, according to the qualitative feedback, that tag recommendations are more useful on average in the class-based interface. Recalling that the ontology-based system recommends a filtered version of the tags suggested by the class-based system (filtered according to the population of the ontology), we can hypothesise that a more comprehensive population of the ontology could lead to more useful tag recommendations. Thus, the improvement of the ontology population process is probably crucial to improve the system in general. One way in which this could be further developed, would be by researching on an automatic system for mapping existing tags to concepts of external knowledge-bases, and by using this information to automatically classify tags in the defined tag categories.

Overall, besides the specific task of tag recommendation, the work presented in this chapter describes an initial approach to a more semantically-driven sound annotation process. The results we report here show that several improvements should be made for deploying such a system in a real-world scenario. However, the inclusion of an ontology in the resource annotation process opens up many possibilities for further researching and improving the system. In the following chapter, we end this dissertation by summarising our work and contributions, and with a discussion about future directions that could be taken to further improve tag recommendation and tagging systems in general (Sec. 7.3).

CHAPTER 7

Summary and future perspectives

7.1 Introduction

In this thesis we have described a number of computational approaches for helping the users of online sharing platforms to better annotate the content they generate. Our approaches are meant to be a step towards increasing the value of resources shared in online sharing platforms by improving their descriptions and enabling better organisation, browsing and searching functionalities. Throughout our thesis, we have contemplated one of the many ways in which the *annotation problem* can be approached. We have focused on the particular task of tag recommendation, for which we advanced its state of the art by proposing novel folksonomy-based recommendation methods and empirically assessing their impact in a real-world sharing platform. In particular, we worked on the case of sound sharing. As explained in Sec. 1.5, sound sharing poses some particularly interesting challenges that highly motivated our research. Nevertheless, we strived for proposing methodologies that can be easily generalised to other multimedia domains.

We started with an introduction to tagging systems, tag recommendation, and the particular case of sound sharing (Chapter 1). We continued by summarising the existing literature on the characterisation of tagging systems and on tag recommendation approaches (Chapter 2). Then, we described and evaluated our first proposed folksonomy-based tag recommendation methods (Chapter 3). We next proposed an improvement over these methods by incorporating some domain-specific knowledge in the form of an audio classifier (Chapter 4), and evaluated the impact of that recommendation method in the real-world tagging system of Freesound (Chapter 5). Finally, and motivated by the findings reported in the previous chapters, we explored a new approach for tag recommendation in which we introduced an audio-specific ontology to inform

the recommendation process and improve in this way the quality of produced annotations (Chapter 6).

In each chapter, we included a section summarising the relevant results and conclusions of the corresponding work. Here, we provide a summary of our contributions from a global point of view (Sec. 7.2). We end this dissertation with a discussion about future research directions (Sec. 7.3), not only related to the particular task of tag recommendation, but also to tagging systems in general.

7.2 Summary of contributions

This thesis contributes to the advancement of the state of the art in tagging systems and, more specifically, tag recommendation and folksonomy-based tag recommendation. The main contributions of this thesis can be summarised as follows:

- It provides a comprehensive overview of tagging systems and discusses about their typical problems and proposed solutions, with the main focus on tag recommendation and the particular case of sound sharing.
- It describes a general scheme for folksonomy-based tag recommendation systems, proposing several alternative strategies for computing each one of the steps of the scheme, as well as comparing the resulting recommendation methods against state of the art approaches. The proposed methods are not only evaluated using a large-scale dataset of audio resources from Freesound, but also using an alternative dataset of similar size composed of image resources from Flickr. Noticeably, the proposed scheme includes a novel step for selecting the number of tags to recommend, which is typically omitted in related research.
- It proposes a successful enhancement to the folksonomy-based tag recommendation scheme by introducing domain-specific knowledge in the form of resource categories that can be automatically detected through a classification step. Noticeably, this tag recommendation system has been deployed in a large-scale and real-world sound sharing platform.
- It explores a new perspective for tag recommendation by proposing a system in which a domain-specific ontology is used to provide tag recommendations. Using this ontology, the system is able to guide the annotation process and, at the same time, it preserves the flexibility of traditional tagging systems.
- It provides a number of methodologies for evaluating tag recommendation systems with and without the intervention of users. Standard information retrieval evaluation methodologies are used to quantitatively asses

several aspects and parameter configurations of the proposed tag recommendation methods. Additionally, user-based evaluations are carried out both in controlled environments and in real-world scenarios to analyse the systems from an empirical point of view.

- It analyses the impact of a tag recommendation system into the folksonomy of a real-world and large-scale sound sharing platform. This analysis includes the definition of a number of metrics and a methodology which are also relevant contributions of this thesis. To the best of our knowledge, this is the first analysis of its kind to be performed in a real-world and large-scale environment.

The research carried out in this thesis has been published in the form of several papers in top international conferences and journals. The outcomes of Chapter 3 have been published in a conference paper and a journal paper (Font & Serra, 2012; Font et al., 2013b). Similarly, the parts of the research presented in Chapter 4 related with the classification step have been published in a conference paper (Font et al., 2014b), and those related with the description and evaluation of the extended tag recommendation method have been published in a journal paper (Font et al., 2014c). Furthermore, the outcome of the research carried out in Chapter 5 has been accepted for publication as a journal paper (Font et al., 2015), and some parts of the definition of the ontology-based recommendation system of Chapter 6 have also been published as a conference paper (Font et al., 2014a). The full list of the author's publications is provided in Appendix B.

7.3 Directions for future research

In the present thesis we have shown several tag recommendation methods which incrementally included more domain-specific knowledge. The approaches we followed have been mainly restricted to the analysis of the folksonomies of tagging systems, and have not included other typical sources of information such as the analysis of resources' content. Even though in Chapter 6 we introduced the use of an ontology to drive the recommendation process, we just started exploring the possibilities of using that ontology and the implications that it might have, not only in tag recommendation, but in tagging systems in general. Hence, we devise two clear perspectives for future research which we now discuss.

Firstly, we believe that the recommendation approaches described in this thesis could be improved with the inclusion of resources' content analysis in the tag recommendation process. On the one hand, using content-based resource classification (e.g., Casey, 2002; Roma et al., 2010) combined with tag-based resource classification for the class-based recommendation method would allow

to predict the resource category before the introduction of the first input tag. This would, for example, allow us to automatically suggest a tag describing that category, or even pre-fill the annotation with that tag. On the other hand, a content-based approach could also be used to select candidate tags based on resource similarity (e.g., Turnbull et al., 2008; Wu et al., 2009), or even by using content-based models to predict tags (e.g., Martínez et al., 2009; Ivanov et al., 2010). Using these strategies, the system would also be able to recommend tags before the introduction of the first input tag, and then combine content-based recommendations with folksonomy-based recommendations in later stages (e.g., Wu et al., 2009; Liu et al., 2010a). In relation to that, the tag categories defined in the ontology could be used as a guideline for defining content-based approaches to recommend tags. For example, content-based models could be built on a tag category basis. Furthermore, this process could be automatically computed by using examples of already annotated resources in the tagging system, and be retrained automatically as new resources were uploaded in the sharing platform.

Secondly, another research direction is the further exploitation of the ontology, not only for the task of tag recommendation, but also as an underlying element in tagging systems and sharing platforms in general. Important challenges in that direction are the definition of comprehensive yet easy to use domain-specific ontologies, and the design of automatic (or semi-automatic) methods for its population. Such ontologies could include more complex and meaningful class hierarchies for resource and tag categories, and be able to represent more meaningful relations among them. To populate ontologies, we believe that approaches for automatically matching tags to concepts of external knowledge bases are a promising direction (e.g., Specia & Motta, 2007; Angeletou, 2008; Moro et al., 2014). Using such mappings and proper disambiguation processes, it would be possible to populate the ontology with a comprehensive set of well-defined tags, thus being able to provide better recommendations. Note that such an approach is close to the idea of using a controlled vocabulary. However, the way in which we envision such systems would allow the flexibility of traditional tagging systems by still allowing the introduction of unknown (i.e., unmatched) tags. The semantic meaning of these unkown tags could, nevertheless, be narrowed down with the usage of tag categories such as we demonstrated in the tagging interface described in Chapter 6.

Another interesting research direction related to the use of tag categories in the annotation process is the evaluation of the capacity of such a tagging system for automatically populating its underlying ontology. Provided that users upload and describe new resources using tag categories, it would be possible to automatically further populate the ontology with previously unseen tags that would be introduced under existing tag categories. From that point of view, it would be interesting to analyse the impact of such a system in the folksonomies emerging from tagging systems, and see how these "semi-structured"

folksonomies could be better exploited for knowledge mining or further ontology refinement (Limpens et al., 2009a). In the same vein, another aspect to evaluate is if the usage of this limited number of tag categories in combination with free-form tags would pose significant limitations for making expressive annotations. In this case, it would be interesting to investigate whether an ontology that could be evolved and edited by users of a tagging platform could be effectively maintained and allow for flexible but structured annotations (e.g., Stojanovic et al., 2002; Braun et al., 2007). A well-populated ontology could also be used to tackle the typical polysemy and synonymy problems found in folksonomies, by explicitly defining these relations among tag instances (e.g., Echarte et al., 2007; Lohmann et al., 2011). In addition, these explicit semantic relations between tags could be also used to provide domain-specific query expansion functionality in the search engines of sharing platforms (Bhogal et al., 2007). For example, user queries could be automatically expanded by including synonym terms taken from the ontology, and results could be grouped in clusters according to alternative meanings of the query terms.

In summary, we believe that the use of content-based strategies to help in resource annotation and the further exploitation of underlying ontologies in tagging systems allows for many improvements in the current functionalities of sharing platforms. For example, searching of content resources could be enhanced by defining complex queries operating over facets corresponding to tag and resource categories, and browsing could also be enhanced by hierarchically organising resources according to these categories. Also, similarity measures for multimedia resources could additionally use the concepts of tag and resource categories to define narrower scopes for the similarity search and, for example, provide complementary similarity scores by treating different tag categories as similarity facets (Bogdanov et al., 2011). Finally, we believe that the use of underlying ontologies, tightly coupled with the annotation systems of sharing platforms, would enable an almost direct publication of resources' metadata as meaningful linked data[47] (Bizer et al., 2009). Overall, such improvements would allow a better exploitation of the huge value of content resources in online sharing platforms. Also, these would represent a step towards conciliating the rate at which user generated content is being created with the ability of computational systems to properly index, organise, and make this information available.

Essentially, for information systems to become more intelligent, they need to better represent and handle knowledge about their domain. By designing new algorithms and ways in which these can take advantage of available data, we will improve our capabilities for sharing information. But perhaps more importantly, we need to focus on understanding and representing that information and its domain, and be able in this way to reason at a level which is presumably closer to how we humans process and share information. For example, a

[47] `http://linkeddata.org`

sound sharing platform with knowledge about musical instruments and genres might allow us to browse instrument sounds in a way that only those relevant for a particular music genre would be displayed. If the platform also embedded knowledge about musical theory, it could also group sounds according to the different musical functions (or roles) that these might play in a composition. Similarly, using this knowledge, such a system could help users in, for example, properly annotating a music loop by suggesting musical genres given some instruments present in the loop. However, to perform these kinds of *reasoning*, an information system should be aware of the related knowledge or *facts*. In this case, such a system should know, for example, that distorted guitars are very prominent in heavy metal music but not in classical music, or that reggae recordings often feature deep bass lines and other harmonic and rhythmic elements playing off beat. These are the kind of systems that we would like to interact with in the future.

> "Now, what *we* want is Facts. Teach these *systems* and *algorithms* nothing but Facts. Facts alone are wanted in life. Plant nothing else, and root out everything else. You can only form the minds of reasoning *computers* upon Facts; nothing else will ever be of any service to them."
>
> Charles Dickens, *Hard Times – For These Times*[48].

[48] Charles Dickens' *Hard Times – For These Times* was first published in 1854. The quote we allude here is found at the opening of the book, and it originally goes as follows: "Now, what I want is Facts. Teach these boys and girls nothing but Facts. Facts alone are wanted in life. Plant nothing else, and root out everything else. You can only form the minds of reasoning animals upon Facts; nothing else will ever be of any service to them.".

Frederic Font Corbera, Barcelona, 11 March 2015.

Bibliography

Adrian, B., Sauermann, L., & Roth-Berghofer, T. (2007). ConTag: A semantic tag recommendation system. In *Proceedings of the International Conference on New Media Technology and Semantic Systems (I-Media and I-Semantics)*, pp. 297–304.

Alstott, J., Bullmore, E., & Plenz, D. (2014). Powerlaw: A python package for analysis of heavy-tailed distributions. *PLoS ONE*, *9*(1).

Ames, M. & Naaman, M. (2007). Why we tag: Motivations for annotation in mobile and online media. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pp. 971–980.

Anderson, A., Ranghunathan, K., & Vogel, A. (2008). Tagez: Flickr tag recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Angeletou, S. (2008). Semantic enrichment of folksonomy tagspaces. In *Proceedings of the International Semantic Web Conference (ISWC)*, pp. 889–894.

Ballan, L., Bertini, M., Del Bimbo, A., Meoni, M., & Serra, G. (2010). Tag suggestion and localization in user-generated videos based on social knowledge. In *Proceedings of the ACM SIGMM Workshop on Social Media (WSM)*, pp. 3–7.

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.

Barrat, A., Barthélemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, *101*(11), 3747–3752.

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 (ICASSP)*, pp. 725–728.

Bennett, K. P. & Campbell, C. (2000). Support vector machines: Hype or hallelujah? *ACM SIGKDD Explorations Newsletter*, *2*(2), 1–13.

Bhogal, J., Macfarlane, A., & Smith, P. (2007). A review of ontology based query expansion. *Information Processing and Management*, *43*(4), 866–886.

Bischoff, K., Firan, C. S., Nejdl, W., & Paiu, R. (2008). Can all tags be used for search? In *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM)*, pp. 203–212.

Bischoff, K., Firan, C. S., Paiu, R., Nejdl, W., Laurier, C., & Sordo, M. (2009). Music mood and theme classification-a hybrid approach. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pp. 657–662.

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. *International Journal on Semantic Web and Information Systems*, *5*(3), 1–22.

Bogdanov, D., Serrà, J., Wack, N., Herrera, P., & Serra, X. (2011). Unifying low-level and high-level music similarity measures. *IEEE Transactions on Multimedia*, *13*(4), 687–701.

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J., & Serra, X. (2013). ESSENTIA: An audio analysis library for music information retrieval. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pp. 493–498.

Braun, S., Schmidt, A., & Walter, A. (2007). Ontology maturing: A collaborative web 2.0 approach to ontology engineering. In *Proceedings of the WWW Workshop on Social and Collaborative Construction of Structured Knowledge (CKC)*.

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the International Conference on World Wide Web (WWW)*, pp. 107–117.

Cano, P., Koppenberger, M., Le Groux, S., Ricard, J., Wack, N., & Herrera, P. (2005). Nearest-neighbor automatic sound annotation with a WordNet taxonomy. *Journal of Intelligent Information Systems*, *24*(2), 99–111.

Cantador, I., Konstas, I., & Jose, J. M. (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 based on supervised ranking model. In *Proceedings of the Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*, pp. 35–48.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, *22*(2), 249–254.

Casey, M. (2002). General sound classification and similarity in MPEG-7. *Organised Sound*, *6*(2), 153–164.

Catt, D. (2006). Advanced tagging and tripletags. `http://revdancatt.com/2006/01/11/advanced-tagging-and-tripletags`. Retrieved: 11 March 2015.

Cattuto, C. (2006). Semiotic dynamics in online social communities. *The European Physical Journal C-Particles and Fields*, *37*(2), 33–37.

Chachada, S. & Kuo, C.-C. J. (2013). Environmental sound recognition: A survey. In *Proceedings of the Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 1–9.

Chen, H.-M., Chang, M.-H., Chang, P.-C., Tien, M.-C., Hsu, W. H., & Wu, J.-L. (2008). SheepDog: Group and tag recommendation for flickr photos by automatic search-based learning. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pp. 737–740.

Chen, L., Wright, P., & Nejdl, W. (2009). Improving music genre classification using collaborative tagging data. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 84–93.

Chen, L., Xu, D., Tsang, I. W., & Luo, J. (2010a). Tag-based web photo retrieval improved by batch mode re-tagging. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3440–3446.

Chen, Z., Cao, J., Song, Y., Guo, J., Zhang, Y., & Li, J. (2010b). Context-oriented web video tag recommendation. In *Proceedings of the International Conference on World Wide Web (WWW)*, pp. 1079–1080.

Cialdini, R. B. (2003). *Influence: Science and practice*. Boston, MA: Allyn and Bacon, 5th edn.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*(4), 661–703.

Corder, G. W. & Foreman, D. I. (2009). *Nonparametric statistics for non-statisticians: A step-by-step approach*. John Willey & Sons.

Cui, J., Wen, F., Xiao, R., Tian, Y., & Tang, X. (2007). EasyAlbum: An interactive photo annotation system based on face clustering and re-ranking. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pp. 367–376.

De Meo, 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.

De Meo, P., Ferrara, E., Abel, F., Aroyo, L., & Houben, G.-J. (2013). Analyzing user behavior across social sharing environments. *ACM Transactions on Intelligent Systems and Technology*, *5*(1).

Ding, Y., Jacob, E. K., Fried, M., Toma, I., Yan, E., Foo, S., & Milojević, S. (2010). Upper tag ontology for integrating social tagging data. *Journal of the American Society for Information Science and Technology*, *61*(3), 505–521.

Doerfel, S. & Jäschke, R. (2013). An analysis of tag-recommender evaluation procedures. In *Proceedings of the ACM Conference on Recommender Systems (RecSys)*, pp. 343–346.

Dunbar, R. (1998). *Grooming, gossip and the evolution of language*. Harvard University Press.

Echarte, F., Astrain, J. J., Córdoba, A., & Villadangos, J. (2007). Ontology of folksonomy: A New modeling method. In *Proceedings of the Semantic Authoring, Annotation and Knowledge Markup Workshop (SAAKM)*.

Fan, J., Shen, Y., Zhou, N., & Gao, Y. (2010). Harvesting large-scale weakly-tagged image databases from the web. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 802–809.

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 (GROUP)*, pp. 351–360.

Feng, S., Lang, C., & Xu, D. (2010). Beyond tag relevance: Integrating visual attention model and multi-instance learning for tag saliency ranking. In *Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR)*, pp. 288–295.

Fichter, D. (2006). Intranet applications for tagging and folksonomies. *Online*, *30*(3), 43–45.

Font, F., Oramas, S., Fazekas, G., & Serra, X. (2014a). Extending tagging ontologies with domain specific knowledge. In *Proceedings of the International Semantic Web Conference (ISWC)*.

Font, F., Roma, G., Herrera, P., & Serra, X. (2012). Characterization of the Freesound online community. In *Proceedings of the International Workshop on Cognitive Information Processing (CIP)*, pp. 279–284.

Font, F., Roma, G., & Serra, X. (2013a). Freesound technical demo. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pp. 411–412.

Font, F., Serrà, J., & Serra, X. (2013b). Folksonomy-based tag recommendation for collaborative tagging systems. *International Journal on Semantic Web and Information Systems*, *9*(2), 1–30.

Font, F., Serrà, J., & Serra, X. (2014b). Audio clip classification using social tags and the effect of tag expansion. In *Proceedings of the AES Conference on Semantic Audio*.

Font, F., Serrà, J., & Serra, X. (2014c). Class-based tag recommendation and user-based evaluation in online audio clip sharing. *Knowledge Based Systems*, *67*, 131–142.

Font, F., Serrà, J., & Serra, X. (2015). Analysis of the impact of a tag recommendation system in a real-world folksonomy. *ACM Transactions on Intelligent Systems and Technology*. In press.

Font, F. & Serra, X. (2012). Analysis of the folksonomy of Freesound. In *Proceedings of the CompMusic Workshop*, pp. 48–54.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, *30*(11), 964–971.

Garg, N. & Weber, I. (2008). Personalized, interactive tag recommendation for Flickr. In *Proceedings of the ACM Conference Recommender Systems (RecSys)*, pp. 67–74.

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

Good, B., Kawas, E., & Wilkinson, M. (2007). Bridging the gap between social tagging and semantic annotation: E.D. the Entity Describer. In *Nature Precedings*.

Gupta, M., Li, R., Yin, Z., & Han, J. (2010). Survey on social tagging techniques. *ACM SIGKDD Explorations Newsletter*, *12*(1), 58–72.

Guy, M. & Tonkin, E. (2006). Folksonomies: Tidying up tags? *D-Lib Magazine*, *12*(1).

Halpin, H., Robu, V., & Shepard, H. (2006). The dynamics and semantics of collaborative tagging. In *Proceedings of the Semantic Authoring and Annotation Workshop (SAAW)*, pp. 1–21.

Herrera, P., Peeters, G., & Dubnov, S. (2003). Automatic classification of musical instrument sounds. *Journal of New Music Research*, *32*(1), 3–21.

Heymann, P. & Garcia-Molina, H. (2006). Collaborative creation of communal hierarchical taxonomies in social tagging systems. Tech. rep., Stanford University.

Hotho, A., Jäschke, R., Schmitz, C., & Stumme, G. (2006). Information retrieval in folksonomies: Search and ranking. In *Proceedings of the European Semantic Web Conference (ESWC)*, pp. 411–426.

Huang, T., Dagli, C., Rajaram, S., Chang, E., Mandel, M., Poliner, G., & Ellis, D. (2008). Active learning for interactive multimedia retrieval. *Proceedings of the IEEE*, *96*(4), 648–667.

Hwang, S.-h. (2007). A triadic approach of hierarchical classes analysis on folksonomy mining. *International Journal of Computer Science and Network Security*, *7*(8), 193–198.

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 ACM International Conference on Multimedia Information Retrieval (MIR)*, pp. 497–506.

Jacob, E. K. (2004). Classification and categorization: A difference that makes a difference. *Library Trends*, *52*(3), 515–540.

Jäschke, R., Eisterlehner, F., Hotho, A., & Stumme, G. (2009). Testing and evaluating tag recommenders in a live system. In *Proceedings of the ACM Conference on Recommender Systems (RecSys)*, pp. 369–372.

Jäschke, R., Hotho, A., Mitzlaff, F., & Stumme, G. (2012). *Challenges in tag recommendations for collaborative tagging systems*, chap. 3, pp. 65–87. Springer Berlin Heidelberg.

Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., & Stumme, G. (2007). Tag recommendations in folksonomies. In *Proceedings of the Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, Lecture Notes in Computer Science, pp. 506–514.

Jeffries, A. (2013). The man behind Flickr on making the service 'awesome again'. `http://www.theverge.com/2013/3/20/4121574/flickr-chief-markus-spiering-talks-photos-and-marissa-mayer`. Retrieved: 11 March 2015.

Kaplan, A. M. & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, *53*(1), 59–68.

Kennedy, L. S., Chang, S.-f., & Kozintsev, I. V. (2006). To search or to label? Predicting the performance of search-based automatic image classifiers. In *Proceedings of the ACM International Workshop on Multimedia Information Retrieval*, pp. 249–258.

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, *54*(3), 241–251.

Kim, H.-L., Breslin, J., Kim, H.-G., & Choi, J.-H. (2010). Social semantic cloud of tags: Semantic model for folksonomies. *Knowledge Management Research and Practice*, *8*(3), 193–202.

Krumm, J., Davies, N., & Narayanaswami, C. (2008). User-generated content. *IEEE Pervasive Computing*, pp. 10–11.

Kuo, C.-C. J. & Zhang, T. (1999). Classification and retrieval of sound effects in audiovisual data management. In *Proceedings of the Asilomar Conference on Signals, Systems, and Computers*, pp. 730–734.

Laurier, C., Grivolla, J., & Herrera, P. (2008). Multimodal music mood classification using audio and lyrics. In *Proceedings of the International Conference on Machine Learning and Applications (ICMLA)*, pp. 688–693.

Lee, S., De Neve, W., Plataniotis, K. N., & Ro, Y. M. (2010). MAP-based image tag recommendation using a visual folksonomy. *Pattern Recognition Letters*, *31*(9), 976–982.

Lessing, L. (2008). *Remix: Making art and commerce thrive in the hybrid economy*. Penguin Press.

Li, J. & Wang, J. Z. (2008). Real-time computerized annotation of pictures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(6), 985–1002.

Li, X., Snoek, C. G. M., & Worring, M. (2009). Learning social tag relevance by neighbor voting. *IEEE Transactions on Multimedia*, *11*(7), 1310–1322.

Limpens, F., Gandon, F. L., & Buffa, M. (2009a). Linking folksonomies and ontologies for supporting knowledge sharing: A state of the art. Tech. rep., Institut National de Recherche en Informatique et Automatique (INRIA).

Limpens, F., Monnin, A., Laniado, D., & Gandon, F. (2009b). NiceTag ontology: Tags as named graphs. In *Proceedings of the Asian Semantic Web Conference (ASWC)*.

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)*, pp. 84–95.

Liu, D., Hua, X.-S., Wang, M., & Zhang, H.-J. (2010a). Image retagging. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pp. 491–500.

Liu, D., Hua, X.-S., Yang, L., Wang, M., & Zhang, H.-J. (2009). Tag ranking. In *Proceedings of the International Conference on World Wide Web (WWW)*, pp. 351–360.

Liu, D., Hua, X. S., & Zhang, H. J. (2011). Content-based tag processing for Internet social images. *Multimedia Tools and Applications*, *51*(2), 723–738.

Liu, X., Yan, S., Luo, J., Tang, J., Huang, Z., & Jin, H. (2010b). Nonparametric label-to-region by search. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3320–3327.

Livshin, A., Peeters, G., & Rodet, X. (2003). Studies and improvements in automatic classification of musical sound samples. In *Proceedings of the International Computer Music Conference (ICMC)*, pp. 171–178.

Lohmann, S., Díaz, P., & Aedo, I. (2011). MUTO: The modular unified tagging ontology. In *Proceedings of the International Conference on Semantic Systems (I-Semantics)*, pp. 95–104.

Lops, P., de Gemmis, M., Semeraro, G., Musto, C., & Narducci, F. (2012). Content-based and collaborative techniques for tag recommendation: an empirical evaluation. *Journal of Intelligent Information Systems*, *40*(1), 41–61.

Macgregor, G. & Mcculloch, E. (2006). Collaborative tagging as a knowledge organisation and resource discovery tool. *Library Review*, *55*(5), 291–300.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge, UK: Cambridge University Press.

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)*, pp. 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 the International Conference on World Wide Web (WWW)*, pp. 641–650.

Marlow, C., Naaman, M., Boyd, D., & Davis, M. (2006). HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, ToRead. In *Proceedings of the ACM Conference on Hypertext and Hypermedia (Hypertext)*, pp. 31–41.

Martínez, E., Celma, O., Sordo, M., Jong, B. D., & Serra, X. (2009). Extending the folksonomies of Freesound using content-based audio analysis. In *Proceedings of the Sound and Music Computing Conference (SMC)*, pp. 23–29.

Mathes, A. (2004). Folksonomies – Cooperative classification and communication through shared metadata. *Computer Mediated Communication*, *LIS590CMC*(Doctoral Seminar).

Merholz, P. (2004). Metadata for the masses. `http://www.adaptivepath.com/ideas/e000361/`. Retrieved: 11 March 2015.

Mika, P. (2007). Ontologies are us: A unified model of social networks and semantics. *Web Semantics: Science, Services and Agents on the World Wide Web*, *5*(1), 5–15.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Moro, A., Raganato, A., Navigli, R., Informatica, D., & Elena, V. R. (2014). Entity Linking meets word sense disambiguation: A unified approach. *Transactions of the Association for Computational Linguistics*, *2*, 231–244.

Naaman, M. & Nair, R. (2008). ZoneTag's collaborative tag suggestions: What is this person doing in my phone? *IEEE MultiMedia*, *15*(3), 34–40.

Newman, R. (2005). Tag ontology design. `http://www.holygoat.co.uk/projects/tags/`. Retrieved: 11 March 2015.

Papadopoulos, S., Kompatsiaris, Y., & Vakali, A. (2010). A graph-based clustering scheme for identifying related tags in folksonomies. In *Proceedings of the International Conference on Data Warehousing and Knowledge Discovery (DaWaK)*, pp. 65–76.

Passant, A. (2007). Using ontologies to strengthen folksonomies and enrich Information retrieval in weblogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.

Passant, A. & Laublet, P. (2008). Meaning of a tag: A collaborative approach to bridge the gap between tagging and linked data. In *Proceedings of the WWW Linked Data on the Web Workshop (LDOW)*.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image and Vision Computing*, *28*(6), 976–990.

Prokofyev, R., Boyarsky, A., Ruchayskiy, O., Aberer, K., Demartini, G., & Cudré-Mauroux, P. (2012). Tag recommendation for large-scale ontology-based information systems. In *Proceedings of the International Semantic Web Conference (ISWC)*, pp. 325–336.

Quintarelli, E. (2005). Folksonomies: Power to the people. `http://www.iskoi.org/doc/folksonomies.htm`. Retrieved: 11 March 2015.

Rae, A., Sigurbjörnsson, B., & Van Zwol, R. (2010). Improving tag recommendation using social networks. In *Proceedings of the Conference on Adaptivity, Personalization and Fusion of Heterogeneous Information (RIAO)*, pp. 92–99.

Rafaeli, S. & Raban, D. R. (2005). Information sharing online: a research challenge. *International Journal of Knowledge and Learning*, *1*(1), 62–79.

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)*, pp. 235–242.

Robu, V., Halpin, H., & Shepherd, H. (2009). Emergence of consensus and shared vocabularies in collaborative tagging systems. *ACM Transactions on the Web*, *3*(4).

Roma, G., Janer, J., Kersten, S., Schirosa, M., Herrera, P., & Serra, X. (2010). Ecological acoustics perspective for content-based retrieval of environmental sounds. *EURASIP Journal on Audio, Speech, and Music Processing*, *2010*(7).

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.

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

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. (2009). *Multivariate density estimation: theory, practice, and visualization.* Probability & Mathematical Statistics. John Willey & Sons.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47.

Sen, S., Lam, S., Rashid, A., & Cosley, D. (2006). Tagging, communities, vocabulary, evolution. In *Proceedings of the Conference on Community Supported Cooperative Work (CSCW)*, pp. 181–190.

Sevil, S. G., Kucuktunc, O., Duygulu, P., & Can, F. (2010). Automatic tag expansion using visual similarity for photo sharing websites. *Multimedia Tools and Applications*, *49*(1), 81–99.

Shirky, C. (2005). Ontology is overrated: Categories, links, and tags. `http://www.shirky.com/writings/ontology_overrated.html`. Retrieved: 11 March 2015.

Sigurbjörnsson, B. & Zwol, R. (2008). Flickr tag recommendation based on collective knowledge. In *Proceedings of the International Conference on World Wide Web (WWW)*, pp. 327–336.

Silvermann, B. W. (1986). *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability. London: Chapman & Hall/CRC.

Simons, J. (2008). Tag-elese or the language of tags. *The Fibreculture Journal*, *FCJ-083*.

Smith, T. (2009). The social media revolution. *International Journal of Market Research*, *51*(4), 559–561.

Solé, R. (2005). Language: syntax for free? *Nature*, *434*(7031).

Song, Y., Zhuang, Z., Li, H., Zhao, Q., Li, J., Lee, W.-c., & Giles, C. L. (2008). Real-time automatic tag recommendation. In *Proceedings of the ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 515–522.

Sood, S. C., Owsley, S. H., Hammond, K. J., & Birnbaum, L. (2007). TagAssist: Automatic tag suggestion for blog posts. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, pp. 1–8.

Sordo, M. (2012). *Semantic annotation of music collections: A computational approach*. Ph.D. thesis, Universitat Pompeu Fabra, Barcelona.

Specia, L. & Motta, E. (2007). Integrating folksonomies with the semantic web. In *Proceedings of the European Semantic Web Conference (ESWC)*, pp. 629–639.

Spiteri, L. F. (2007). The structure and form of folksonomy tags: The road to the public library catalog. *Information Technology and Libraries*, *26*(3), 13–25.

Stanoevska-Slabeva, K. (2002). Toward a community-oriented design of internet platforms. *International Journal of Electronic Commerce*, *6*(3), 71–95.

Stojanovic, L., Maedche, A., Motik, B., & Stojanovic, N. (2002). User-driven ontology evolution management. In *Proceedings of the International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, pp. 285–300.

Suh, B. & Bederson, B. (2004). Semi-automatic image annotation using event and torso identification. Tech. rep., Human Computer Interaction Laboratory, University of Maryland.

Sundaram, S. & Narayanan, S. (2008). Classification of sound clips by two schemes: Using onomatopoeia and semantic labels. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1341–1344.

Tang, J., Chen, Q., Yan, S., Chua, T.-S., & Jain, R. (2010). One person labels one million images. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pp. 1019–1022.

Tao, R., Li, Z., & Ji, Y. (2010). Music genre classification using temporal information and support vector machine. In *Proceedings of the Advanced School for Computing and Imaging Conference (ASCI)*.

The YouTube Team (2013). Here's to eight great years. `http://youtube-global.blogspot.com/2013/05/heres-to-eight-great-years.html`. Retrieved: 11 March 2015.

Tian, Y., Liu, W., Xiao, R., Wen, F., & Tang, X. (2007). A face annotation framework with partial clustering and interactive labeling. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8.

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 Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3447–3454.

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.

Ulges, A., Schulze, C., Keysers, D., & Breuel, T. (2008). Identifying relevant frames in weakly labeled videos for training concept detectors. In *Proceedings of the Conference on Image and Video Retrieval (CIVR)*, pp. 9–16.

Vander Wal, T. (2005). Explaining and showing broad and narrow folksonomies. http://www.vanderwal.net/random/entrysel.php?blog=1635. Retrieved: 11 March 2015.

Vander Wal, T. (2007). Folksonomy. http://vanderwal.net/folksonomy.html. Retrieved: 11 March 2015.

Von Ahn, L. (2006). Games with a purpose. *Computer*, *39*(6), 92–94.

Von Ahn, L. & Dabbish, L. (2004). Labeling images with a computer game. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pp. 319 – 326.

Wagner, C., Strohmaier, M., & Huberman, B. (2014). Semantic stability and implicit consensus in social tagging streams. In *Proceedings of the International Conference on World Wide Web (WWW)*, pp. 735–746.

Wahlforss, A. L. & Eric (2013). SoundCloud is 5! http://blog.soundcloud.com/2013/11/13/soundcloud-is-5/. Retrieved: 11 March 2015.

Wang, M. & Hua, X.-S. (2011). Active learning in multimedia annotation and retrieval: A survey. *ACM Transactions on Intelligent Systems and Technology*, *2*(2), 1–21.

Wang, M., Ni, B., Hua, X.-S., & Chua, T.-S. (2012). Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. *ACM Computing Surveys*, *44*(4), 1–24.

Wikipedia (2014a). Remix culture. http://en.wikipedia.org/w/index.php?title=Remix_culture&oldid=614815119. Retrieved: 11 March 2015.

Wikipedia (2014b). Tag - metadata. http://en.wikipedia.org/wiki/Tag_(metadata). Retrieved: 11 March 2015.

Wu, H., Zubair, M., & Maly, K. (2006). Harvesting social knowledge from folksonomies. In *Proceedings of the ACM Conference on Hypertext and Hypermedia (Hypertext)*, pp. 111–114.

Wu, L., Yang, L., Yu, N., & Hua, X.-S. (2009). Learning to tag. In *Proceedings of the International Conference on World Wide Web (WWW)*, pp. 361–370.

Xu, Z., Fu, Y., Mao, J., & Su, D. (2006). Towards the semantic web: Collaborative tag suggestions. In *Proceedings of the WWW Collaborative Web Tagging Workshop*.

Zangerle, E., Gassler, W., & Specht, G. (2011). Using tag recommendations to homogenize folksonomies in microblogging environments. In *Proceedings of the International Conference on Social Informatics (SocInfo)*, pp. 113–126.

Zhang, N., Zhang, Y., & Tang, J. (2009). A tag recommendation system for folksonomy. In *Proceeding of the ACM Workshop on Social Web Search and Mining (SWSM)*, pp. 9–16.

Zhao, W. L., Wu, X., & Ngo, C. W. (2010). On the annotation of web videos by efficient near-duplicate search. *IEEE Transactions on Multimedia*, *12*(5), 448–461.

Appendix A: Freesound

Introduction

Freesound is an online sharing platform where people with diverse interests share recorded sound samples under Creative Commons licenses (Fig. 1). The audio content that can be shared in Freesound is restricted to *sounds*, which may include any kind of audio material like sound effects, environmental recordings or even building blocks for musical compositions, but not music tracks in the traditional sense of "finished" compositions or songs. It was started in 2005 at the Music Technology Group[49] of Universitat Pompeu Fabra, and it is being maintained to support diverse research projects and as a service to the overall research and artistic community. Freesound's initial goal was to give support to sound researchers, who often have trouble finding large royalty-free sound databases to test their algorithms, and to sound artists, who use pre-recorded sounds in their pieces. After eight years since its inception, Freesound has become one of the most popular sites for sharing sound samples, with an average of 45,000 unique visitors per day. More importantly, there is a highly engaged community of users continuously contributing to the site, not only uploading sounds but also commenting, rating and discussing in the forums about relevant topics for the community. All sounds in Freesound are manually moderated by a group of Freesound users (the Freesound moderators) that check for the accurateness of sound annotations and for adequacy of uploaded sounds.

All the content in Freesound is released under Creative Commons licenses. When uploading sounds, Freesound users can choose between CC0 (public domain), Attribution and Attribution-NonCommercial licences[50]. The reason to offer these licenses is to ensure that all the content uploaded in Freesound can be reused by other users, developers and researchers, but at the same time we provide users the option to require the attribution of their work or to restrict the use of their sounds to non-commercial activities. Furthermore, the source code of the web application is released as open source[51] under the GNU AGPL license[52].

Freesound was built with high load and scalability in mind. Fig. 2 shows a block diagram of its architecture. Retrieval of sounds can be performed using text queries, audio content-based similarity search, or by browsing tags or

[49] `http://mtg.upf.edu`
[50] `http://www.creativecommons.org/licenses`
[51] `http://www.github.com/MTG/freesound`
[52] `http://www.gnu.org/licenses/agpl.html`

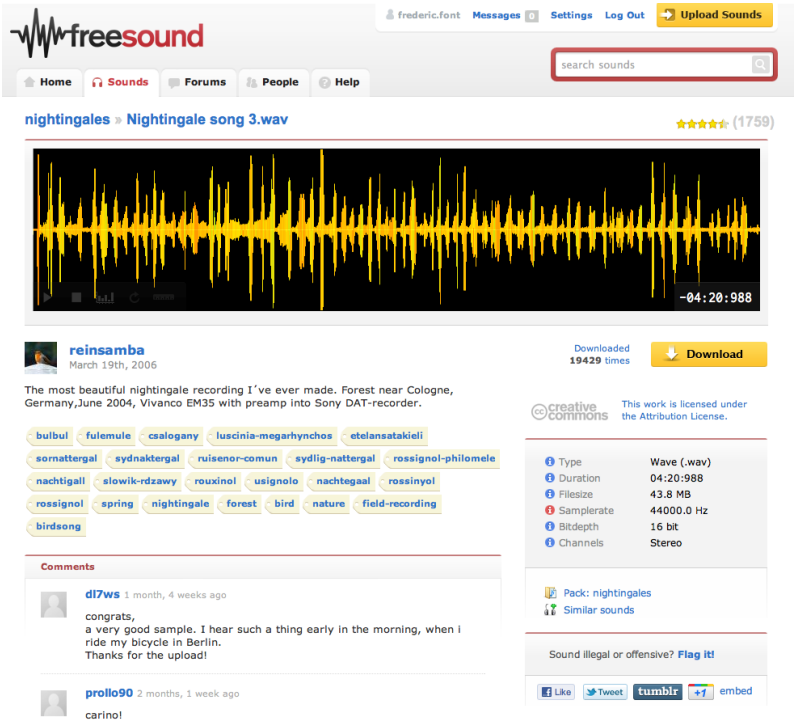


Figure 1: Screenshot of a sound page of Freesound.

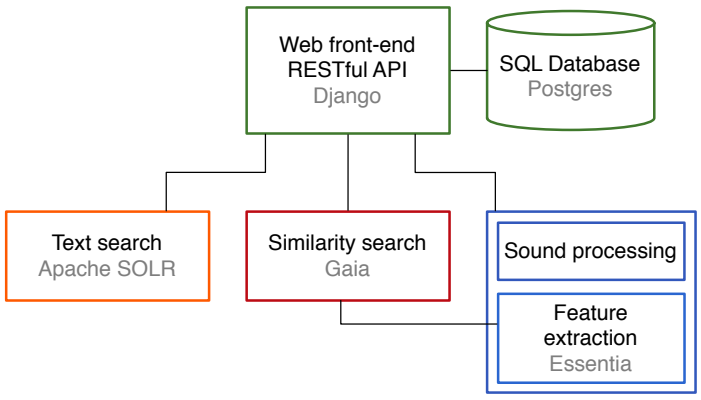


Figure 2: Simplified diagram of the Freesound architecture (Font et al., 2013a).

geotags. Content-based similarity search is performed over a broad set of low-level audio features[53], extracted with Essentia[54] (Bogdanov et al., 2013), an open-source audio feature extraction tool also developed at the Music Technology Group, and indexed in Gaia[55], another open-source tool developed at the Music Technology Group to build and query large feature spaces. The front-end is a Django[56] application which includes basic social interaction features (forum, sound comments, sound ratings, private messaging, etc.), and using a PostgreSQL[57] database for permanent storage. Text indexing is supported by an Apache Solr[58] server including text descriptions and tags, which allows for sophisticated text queries using the Solr query syntax. A distributed architecture is used for processing incoming sounds, producing compressed previews and waveform/spectrogram images, as well as for audio feature extraction. Frame-level and sound-level features are available for each sound.

In 2011, a major update to Freesound was deployed which included a complete redesign of both the backend and the frontend the site, and introduced an API to facilitate access to the Freesound content to researchers and developers[59]. The API runs as a Django application based on the RESTful principles. With the Freesound API users can browse, search, and retrieve information about Freesound users, packs, and the sounds themselves, and also upload, comment, rate and bookmark sounds. Furthermore, the API allows to search for similar sounds to a given target (based on audio content features) and to retrieve content features extracted from the audio files, as well as to perform advanced queries combining content analysis features and other metadata such as tags and textual descriptions.

In the following sections we briefly describe some information about Freeound which can be of interest to the reader of this thesis. First, we provide statistics about some aspects of general interest. Then, we provide further insight into the community of users around Freesound. For further information we refer the reader to previous publications by the authors (Font & Serra, 2012; Font et al., 2012, 2013a).

General numbers

In Table 1 we report some numbers of general interest about several Freesound aspects such as the number of sounds, users, tags and the distinct social activities. Figs. 3, 4 and 5 complement these numbers by showing the evolution of the

[53] A list of these features can be found here: `http://www.freesound.org/apiv2/descriptors/` (requires Freesound account).

[54] `http://essentia.upf.edu`

[55] `http://github.com/mtg/gaia`

[56] `http://www.djangoproject.com`

[57] `http://www.postgresql.org`

[58] `http://lucene.apache.org/solr`

[59] `http://www.freesound.org/docs/api/`

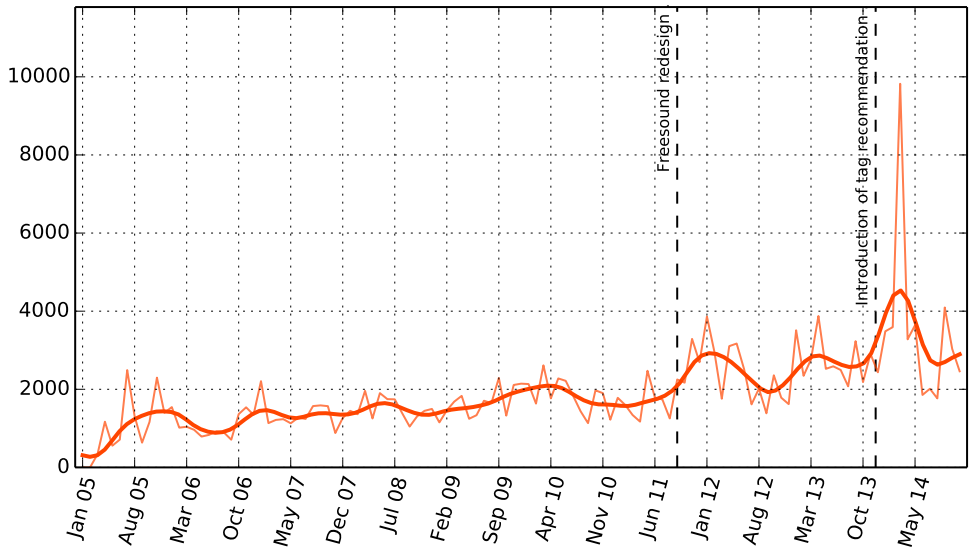


Figure 3: Evolution of the number of uploaded sounds per month. The stronger line corresponds to a smoothed version of the number of uploaded sounds, using a Hann window of 11 points.

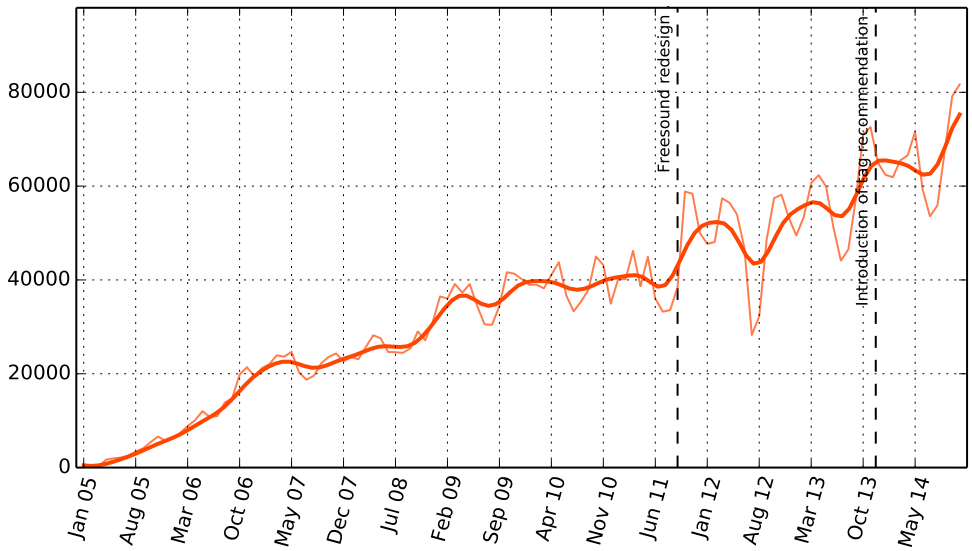


Figure 4: Evolution of the number of newly registered users per month. The stronger line corresponds to a smoothed version of the number of newly registered users, using a Hann window of 11 points.

Number of sounds	230,327	Number of contributor users[a]	14,353
Number of sound packs	14,004	Number of unique tags[b]	77,753
Number of sound comments	191,556	Number of tag applications	1,670,159
Number of sound ratings	929,380	Average number of tags per sound	7.19
Number of sound downloads	65,399,428	Number of forum posts	47,350
Number of registered users	4,341,738	Number of forum threads	9,648

[a] Users that have contributed by uploading, at least, one sound.
[b] Not necessarily semantically unique.

Table 1: Basic statistics of Freesound (at the time of this writing).

number of uploaded sounds, newly registered users, and new tags per month. It is particularly interesting to observe that both the number of uploaded sounds and newly registered users per month has been steadily increasing since the start of Freesound in 2005. Furthermore, it is interesting to note the sudden increase in the number of new tags per month that happened along with the aforementioned redesign of Freesound (Sec. 7.3) and the later decrease after the introduction of the tag recommendation system (Fig. 5). This evolution suggests that the new Freesound design had a huge impact on the way users annotate sounds, resulting in less reuse of tags. This might be explained because, before the redesign, Freesound's annotation interface included a section in which the most commonly used tags by someone annotating a sound were shown. This section was removed after the redesign. After the introduction of tag recommendation however, the rate at which new tags are created diminishes, as we already observed and discussed in Chapter 5 (Sec. 5.3.1) and can also be seen in Fig. 5.

Freesound's community of users

The active community of users behind Freesound is the clearest indication of its success. As we have already seen, the community has been growing over time, reaching more than 4 million registered users and 14,000 unique sound contributors at the time of this writing. Freesound's community can be characterised as a "task-oriented" community, that is to say, a community where its members pursue some collective goals that benefit the whole community (Stanoevska-Slabeva, 2002). To get some insight in that aspect, we carried out a small online survey in the Freesound forums, asking users about their opinion on the existence of shared goals and, if so, which are these goals. A total of 86 Freesound users participated in the survey, 50 of them agreeing with the existence of shared goals, and the others either not directly answering the question (31) or denying the existence of these goals (5). Shared goals that users described in their responses are quite diverse. However, the most repeated goals could be summarized as "sharing sounds" (mentioned by 43% of those participants

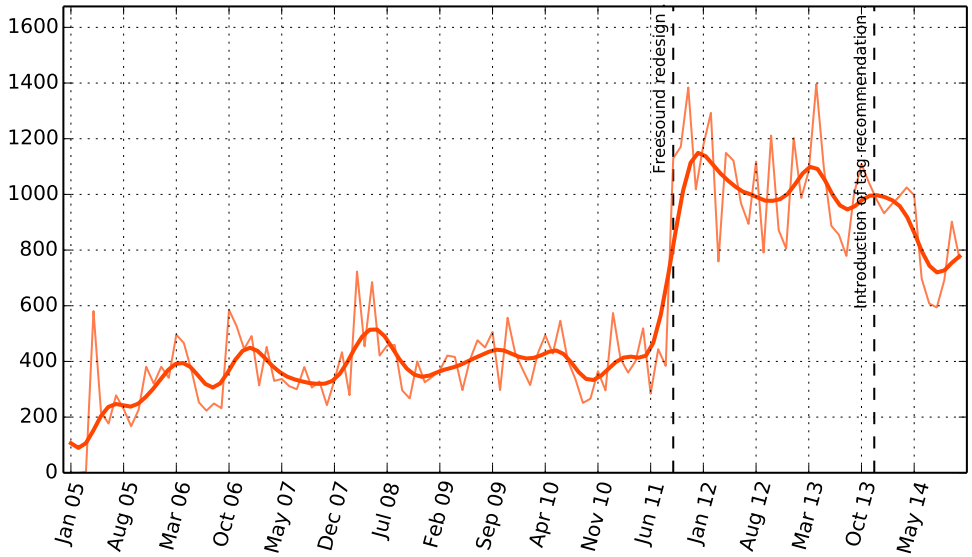


Figure 5: Evolution of the number of new tags per month. The stronger line corresponds to a smoothed version of the number of new tags, using a Hann window of 11 points.

agreeing with the existence of shared goals), "building a big sound archive" (30%) and "helping each other by uploading useful sounds" (21%).

Furthermore, in that same survey we also asked users about the things that make Freesound different from other similar sites. In that case, 66% of users pointed either at the quantity, quality, diversity or "freeness" of accessible sounds, all of them being primary design criteria for Freesound. Other common answers are related to the user interface or the focus on sharing sound samples rather than music (24% of the responses).

Finally, we asked users about the applications for which Freesound is being used (i.e., for what purposes Freesound samples are being reused). Responses show that the most important usage of Freesound samples is in movies and animations (35%), followed by music (20%), theatre (9%), sound design (9%), education and academy (6%), and videogames (5%). Particularly interesting is the fact that the remaining 16% of users pointed out Freesound itself as an application, and hence mainly using Freesound for listening, sharing and contributing sounds, and for its basic social functionalities.

Appendix B: publications by the author

In press

Font, F., Serrà, J., & Serra, X. (2015). Analysis of the impact of a tag recommendation system in a real-world folksonomy. *ACM Transactions on Intelligent Systems and Technology.*

Journal papers

Font, F., Serrà, J., & Serra, X. (2014c). Class-based tag recommendation and user-based evaluation in online audio clip sharing. *Knowledge Based Systems, 67*, 131–142.

Font, F., Serrà, J., & Serra, X. (2013). Folksonomy-based tag recommendation for collaborative tagging systems. *International Journal on Semantic Web and Information Systems, 9*(2), 1–30.

Roma, G., Herrera, P., Zanin, M., Toral, S.L., Font, F., & Serra, X. (2012) Small world networks and creativity in audio clip sharing. *Journal of Social Network Mining, 1*(1), 112–127.

Conference papers

Font, F., Oramas, S., Fazekas, G., & Serra, X. (2014a). Extending Tagging Ontologies with Domain Specific Knowledge. In *Proceedings of the International Semantic Web Conference (ISWC, Poster track).*

Font, F., Serrà, J., & Serra, X. (2014b). Audio clip classification using social tags and the effect of tag expansion. In *Proceedings of the AES Conference on Semantic Audio.*

Font, F., Roma G., & Serra, X. (2013). Freesound Technical Demo. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pp. 411—412.

Font, F., Serrà, J., & Serra, X. (2012). Folksonomy-based tag recommendation for online audio clip sharing. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR 2012)*, pp. 73—78.

Font, F., & Serra, X. (2012). Analysis of the Folksonomy of Freesound. *Proceedings of the CompMusic Workshop*, pp. 48–54.

Font, F., Roma G., Herrera P., & Serra, X. (2012). Characterization of the Freesound Online Community. In *Proceedings of the International Workshop on Cognitive Information Processing (CIP)*, pp. 279–284.

Font, F., & Serra, X. (2011). Extending Sound Sample Descriptions through the Extraction of Community Knowledge. In *Proceedings of the International Conference on User Modeling, Adaptation and Personalization (UMAP)*, pp. 418–421.

Akkermans, V., Font F., Funollet J., De Jong, B, Roma G., Togias S., & Serra, X. (2011). Freesound 2: An Improved Platform for Sharing Audio Clips. Demo session at the *International Conference on Music Information Retrieval (ISMIR)*.