Tag Cloud Reorganization: Finding Groups of Related Tags on Delicious

ABSTRACT

Tag clouds have become an appealing way of navigating through web pages on social tagging systems. Recent research has focused on finding relations among tags to improve visualization and access to web documents from tag clouds. Reorganizing tag clouds according to tag relatedness has been suggested as an effective solution to ease navigation. Most of the approaches either rely on co-occurrences or rely on textual content to represent tags. In this chapter we will explore tag cloud reorganization based on both of them. We compare these clouds from a qualitative point of view, analyzing pros and cons of each approach. We show encouraging results suggesting that co-occurrences produce more compelling reorganization of tag clouds than textual content, being computationally less expensive.

INTRODUCTION

Social bookmarking sites allow to collaboratively annotate web pages, relying on the social tagging philosophy. Social tagging is a Web 2.0 application based phenomenon where users can describe web contents by adding tags or keywords as metadata in an open, collaborative and non-hierarchical way (Smith, 2008). Social bookmarking is a popular way to store, organize, comment on and search links to any web page, and it has emerged as one of the most important web applications that eases information sharing. Popular collaborative tagging sites have aggregated a vast amount of user-created metadata in the form of tags, providing valuable information about the interests and expertise of the users. Because of this, it becomes a fertile area to scientific research on social media (Gupta et al., 2010).

In order to facilitate access to tagged resources, and to enable visual browsing, social bookmarking tools typically provide an interface model known as the tag cloud. A tag cloud is an appealing way to enable users to navigate through the most popular tags of a social bookmarking site. When users access the information contained in these structures, it is presented in the form of a cloud consisting of the most popular tags, where the bigger is the font size of a tag, the more popular it is on the site. Typical tag clouds include between 100 and 200 tags, and tag weights are represented by different font sizes, or other visual clues. In addition, tags can be sorted in alphabetical, size-based, or random order, and users can sometimes customize clouds with different fonts, layouts, and color schemes. These structures are particularly useful for browsing and for information discovery, because they provide a visual summary of the content in the collection. However, related tags do not appear in nearby spaces of the tag cloud, and it is not easy to find the tags of one’s interest. To solve this problem, research in the field has pointed out that grouping related tags, and showing them close to each other can help enhance navigation through tag clouds.

In order to enhance browsing phase in a tag cloud, an effective way is to identify inter-related tags and relations among contents. This book chapter aims to discuss the tag grouping task so that it enables an enhanced visualization and improved navigation through the tag cloud. To this end, several methods of representing tags have been proposed in earlier research. Most of them consider co-occurrences among tags to group related tags into clusters, but do not pay special attention on the algorithm employed to weight such co-occurrences. In this work, we focus on the reorganization of a tag cloud based on the identification of groups of inter-related tags, and compare different methods for weighting tag co-occurrences. We rely on a well-known clustering algorithm for this purpose.

Recently, there has been an increasing interest on tag clustering tasks; most of them tackle the problem from the point of view of tag co-occurrences (Specia and Motta (2007), Mika (2007), Sbodio and Simpson (2009)). Other works have followed a content-based approach, such as Zubiaga et al. (2009). All of them performed a qualitative evaluation of their results, finding appealing groupings for human users. Nevertheless, these works did not compare content-based methods with those based on tag co-occurrences, widely used in the literature.

In this book chapter we further explore several state-of-the-art weighting functions to represent co-occurrences among tags. After clustering tags with these weightings, we compare the results with those obtained by the content-based approach. Going further, we analyze and discuss the appropriateness and performance of each approach.

Next, in Section Background we cover some basic ideas about social tagging systems and present the related work. In Section Tag Cloud Reorganization, we explain the settings of our experiments, including dataset, tag representation approaches and tag clustering algorithm. To conclude the section, we analyze the resulting clouds and discuss some possible applications. Finally, we summarize the future research directions and conclusion.

BACKGROUND

Tagging is an open way to assign tags or keywords to resources or items (e.g., a web page), in order to describe the characteristics of them. This enables later retrieval of resources in an easier way. As opposed to a classical taxonomy-based categorization system, they are usually non-hierarchical, and the vocabulary is open, so it tends to grow indefinitely. For instance, a user could tag this chapter as *social-tagging*, *clustering* and *delicious* whereas another user could use *chapter*, *research* and *tagging* tags to annotate it.

A tagging system becomes social when tags are publicly visible, and so profitable for anyone. The fact of a tagging system being social implies that a user could take advantage of tags defined by others to retrieve a resource. As a result, the collection of tags defined by the community creates a tag-based organization, so-called folksonomy. A folksonomy is also known as a community-based taxonomy, where the classification scheme is plain, there are no predefined tags, and therefore users can freely choose new words as tags.

Depending how users assign tags, two types of social tagging systems can be distinguished (Smith, 2008):

- **Simple Tagging**: users describe their own resources or items, such as photos on Flickr.com, news on Digg.com or videos on Youtube.com, but nobody else tags another user's resources. Usually, the author of the resource is who tags it. This means no more than one user tags a concrete item.
- **Collaborative Tagging**: many users tag the same item, and every person can tag it with his own tags in his own vocabulary. As a result, several users tend to post the same item. For instance, CiteULike.org, LibraryThing.com and Delicious.com are based on collaborative tagging, where each resource (papers, books and URLs, respectively) could be tagged by all the users who considered it interesting.

Figure 1 shows an illustrative example of how each of these systems work.

Figure 1. Simple Tagging vs. Collaborative Tagging

Within the collaborative tagging systems, this work is focused on social bookmarking. In social bookmarking sites, people can post and tag their favorite web pages by using the tags they consider representative. These tags represent the keywords a user would use to look for it, and tend to differ from user to user. Thus, the more users describe an item, the more precise its tag set is. This is one of the main hypotheses of social tagging.

Formally, in a social bookmarking site each user u_i can post an item i_j with a set of tags $T_{ij} = \{t_1,...,t_p\}$, with a variable number p of tags. After k users posted i_j, it is described as a weighted set of tags $T_j = \{w_1 t_1,...,w_n t_n\}$, where $w_1,...,w_n \leq k$.

The nature of social bookmarking sites offers us multiple positive aspects:

- Collaborative tagging of the same item by different users allows to create a weighted list of tags established by general consent. This leads to a wide set of tags, where a few tags are high-weighted and many of them are low-weighted, following a power law distribution.
- The open vocabulary allows users to create non-existing tags required by the current affairs or personal needs.

As far as the research on social tagging is concerned, there are some unhanded aspects that interfere in the performance of social bookmarking sites:

- Different tags can be synonymous (e.g., photo and photography), or have similar or related meanings.
- Different hypernym/hyponym relations can be found (e.g., programming and java), with different levels of specificity for related tags.
- No distinction is made for polysemous tags, such as *library*, which could mean both a place containing books and a collection of sub-programs, amongst others.
- The purpose of users is different when tagging a page from Youtube or a tutorial about movie editing by using a common tag like *video*.

As regards to the nature and purpose of tags, different classification schemes have been proposed. In Golder & Huberman (2006), the authors consider that tags may be categorized in a list of 7 types: (1) identifying what (or who) it is about, (2) identifying what it is, (3) identifying who owns it, (4) refining categories, (5) identifying qualities or characteristics, (6) self reference, and (7) task organizing. These 7 types may be reduced into the following three more general types, according to Sen et al. (2006):

- **Factual tags**: describes item topics, kinds of item or category refinements being objective tags, e.g., *design, video.*
- **Subjective tags**: describes item quality. e.g., *cool, interesting*.
- **Personal tags**: describes item ownership, self-reference or tasks organization. e.g., *toread.*

Even though these classifications of tags include most of the cases, it is not so easy to classify all of them, as the open vocabulary allows to assign new kinds of them. For instance, we can also find **temporal tags**, such as *2008*, which are not considered in the above lists.

In summary, social bookmarking sites allow us to get large annotated datasets, but some kind of preprocessing could be required in order to access the needed information in a suitable way.

Related Work

With the emergence and popularity of social tagging systems, several researchers have shown their concern about improving the navigation through tag clouds. The lack of meaningful spatial interpretations in tag clouds has already been addressed by several authors. Lately, there has been an increasing interest on the discovery of semantic relations among social tags Garcia-Silva et al. (2011). With the aim of getting a reorganized tag cloud, a representation of tags must be performed first. Many of the works represent tags consider co-occurrences among tags, whereas a few rely on the textual content of the tagged documents.

For instance, Dattolo et al. (2011) present an approach to discover tag semantics using clustering techniques to find different categories of related tags. They present an approach that considers distributional measures of tags, besides intersection (co-occurrences) and Jaccard (normalized co-occurrences). They perform a qualitative evaluation over a reduced set of top 20 tags, a group of tags known to be ambiguous, and a set of subjective tags. Vandic et al. (2011) propose a method to improve search on social tagging systems by clustering syntactic variations of tags with the same meaning. They use the cosine similarity based on co-occurrence vectors for measuring semantic relatedness. In a similar approach, Specia & Motta (2007) perform the clustering process based on the similarity among tags given by their co-occurrence, where each tag is represented using the intersection with each other tag in the whole tag set.

In Begelman et al. (2006), they build an undirected graph representing the tag space, where the vertices correspond to tags, and the edges between them represent their co-occurrence frequency. The tag space is built with the pairs of tags that co-occur more frequently than expected, by looking for a cut-off point above which a pair of tags is considered strongly related. The authors obtain clusters of related tags using a clustering algorithm based on the spectral bisection.

Textual content has also been considered to represent and find inter-related tags. In Brooks & Montanez (2006) the authors analyzed document similarity based on weighted word frequency using the TF-IDF term weighting function. They grouped documents with sharing tags into clusters, and then compared the similarity of all documents within a cluster, by means of the average pairwise cosine similarity and an agglomerative algorithm. Zubiaga et al. (2009) present a methodology to obtain and visualize a cloud of related tags based on the use of Self-Organizing Maps, where relations among tags are established taking into account the textual content of the tagged documents. Although the resultant tag cloud was promising, they did not compare the content based representation with any other co-occurrence based representation.

In this context, in addition to other clustering algorithms, Self-Organizing Maps have been used to cluster related tags. An advantage of SOMs over other methods is that the clustering step itself produces a graphical map of the folksonomy. Graph-based clustering methods such as that by Simpson (2008) have also been used to produce a visual graph of tags, but these graphs are often more complex, with many edges, and require more expensive layout algorithms. The visualization capabilities of SOMs provide an intuitive way of representing the distribution of data as well as the object similarities. In Sbodio (2009) tags were clustered by means of a SOM, using a tag representation based on co-occurrence among tags. Once the map was trained, the authors used it to classify new tagged documents. Other works that use Self Organizing Maps to find related tags are Li & Zhu (2008) and Gabrielsson & Gabrielsson (2006).

There are works in clustering tags on the social bookmarking site Last.fm. In Chen et al. (2009) a clustered tag cloud for the social bookmarking site Last.fm − a popular social bookmarking site, where music-related resources like artists and songs can be tagged by users − is presented, and Lehwark et al.

(2008) use a SOM and U-Map techniques to visualize and cluster tagged music data from Last.fm. These two works rely on tagging patterns to discover relations among them, e.g., they group tags containing the string 'rock'. Afterwards, they calculate semantic similarities between tags by means of co-occurrences and other measures like the TF-IDF weighting function, euclidean distance and cosine similarity. Nonetheless, their approaches are not automated but manual, and they manually define similar words to be clustered, so their systems do not allow to easily update the clustered tag cloud.

Different from those above, we perform a qualitative comparison of two tag representations: tag co-occurrences and document textual content.

TAG CLOUD REORGANIZATION

Dataset

First of all, we need a dataset to conduct the experiments. We decided to use the DeliciousT140 dataset released by Zubiaga et al. (2009). This dataset is made up by 144,574 unique URLs, all of them with their corresponding social tags retrieved from Delicious on June 2008. This set of documents is annotated with 67,104 different tags.

Going into further details about this collection, it was created starting from the 140 most popular tags of the site, that is, the whole tag cloud (in the following T140). Each URL is attached to an amount *k* of annotators, and a list of weighted tags $T = \{w_1\ t_1,...,w_n\ t_n\}$, where *n* is at most 25, limited by the social bookmarking site at the time of the dataset generation. Along with the social tags from Delicious, it includes the HTML content corresponding to that URLs. Moreover, the dataset contains only English-written documents. This dataset is available on the web for research purposes.

Tag Representations

The great majority of approaches to represent tags are based on co-occurrences among tags. As far as we know, there is not any comparison between the co-occurrence representation and any other representation based on the textual content of the annotated web documents. At the heart of both approaches is the same kind of information, but they stress in a different way. On the one hand, both take into account the document content, one in an explicit way (content based) and the other in an implicit way (co-occurrence based), since considering tag co-occurrences assumes relations among contents from the tagged documents. On the other hand, both use tag co-occurrence data: one in an explicit way (co-occurrence based) and the other in an implicit way (content based), since the content of a document can take part in the representation of more than one tag we take into account co-occurrence information in an implicit way.

In this chapter, we tried these two approaches described above to represent tags in order to reorganize a tag cloud: representation by tag co-occurrence, and by textual document content. In both cases we use the VSM.

Representation by Tag Co-occurrence

User posts present interesting features to represent tags. When a user tags a document, the implicit semantics of the tag is assigned to document content. Since we considered only popular tags (only the 140 tags in the tag cloud are taken into account) we can expect a reasonable user agreement, and these tags will fit the documents they represent quite well. So, we are taking into account information provided by user classification, in such a way that we could say we are building a tag representation based on human knowledge. Moreover, this classification was performed by a large number of users. Therefore, if we find two highly posted tags labeling the same document, we can assume the document content is related to both of them. Thus, if this tag co-occurrence is found in several documents, being the number of documents large enough to be representative in our dataset, we can conclude that a relation between those tags exists, due to the fact that system users posted the same documents with both of them. From this assumption we formulate our main hypothesis for co-occurrence based tag representation: *the greater the number of documents tagged by the same tags, the greater is the similarity among these tags*.

Based on these ideas, we propose four different tag weighting functions. For each tag we build a vector representing its degree of co-occurrence with every tag within T140. Therefore, we obtained 140 vectors with 140 dimensions each, one per tag. Hence, each vector component corresponds to a different T140 tag, and the value set for this component, hereafter tag weight, measures the degree of co-occurrence between the tag corresponding to that component and the tag represented by the vector. Equation (1.1) shows how a tag vector is organized:

$$Tag_i = (W_{tag_i,tag_1}, \cdots, W_{tag_i,tag_j}, \cdots, W_{tag_i,tag_{140}}) \quad \forall \quad tag_i \in T140 \tag{1.1}$$

being Tag_i the vector representation of tag *i*, and W_{tag_i,tag_j} the weight between Tag_i and Tag_j.

Thus, the tag vectors corresponding to the whole collection (140 vectors) make up the matrix (1.2).

$$\begin{pmatrix} W_{tag_1,tag_1} & \cdots & W_{tag_1,tag_{140}} \\ \vdots & \ddots & \vdots \\ W_{tag_{140},tag_1} & \cdots & W_{tag_{140},tag_{140}} \end{pmatrix} \tag{1.2}$$

So far, we have defined the vector space used to represent tags. We have also talked about the weighting functions used to build the vectors and the main ideas we took into account to choose them. Now, we will define in detail each of the weighting functions. We consider three main features to be combined with the number of documents tagged with both tags: (i) the minimum document frequency between tags, (ii) the maximum tag document frequency between tags, and (iii) the number of documents tagged with at least one of the tags. We combine these 3 weights to define 4 different weighting functions:

- **Document frequency of the intersection of two tags** [Equation (1.3)]: the absolute number of documents in the dataset tagged with both tags. In this case, we make use of the main hypothesis previously formulated directly. This function is not normalized to the dataset dimension and so, its values will not be relative but absolute within the dataset.

$$W_{tag_i,tag_j} = df(tag_i \cap tag_j) \tag{1.3}$$

- **Document frequency of the intersection of two tags over document frequency of the union of those tags** [Equation (1.4)]: this function represents the Jaccard similarity coefficient. If two tags have a high Jaccard score, then they almost always occur in the dataset as a pair, and one will almost never occur in the absence of the other. This function also assumes the main hypothesis, but in this case, the values are scaled-down by the number of documents tagged with one of the tags.

$$W^{union}_{tag_i,tag_j} = \frac{df(tag_i \cap tag_j)}{df(tag_i \cup tag_j)} \tag{1.4}$$

The Jaccard similarity coefficient has been assumed by several tag clustering studies like Simpson (2008) and Sbodio (2009), reason why we consider this tag weighting function within our baseline. However, its appropriateness as compared to other measures has not yet been shown. In this work, we also aim to show whether or not Jaccard is suitable for the task. We consider this tag weighting function within our baseline because it is one of the most used function in the literature.

- **Document frequency of the intersection of two tags over the minimum tag document frequency between them** [Equation (1.5)]: in this case we adjust the value using the minimum tag document frequency of both tags in the dataset, in such a way that the greater the number of documents tagged with the least common tag in connection with the intersection value, the lower the weight is. This function also assumes the previous hypothesis, but in this case, the values are scaled-down by the number of documents tagged with the least common tag.

$$W^{min}_{tag_i,tag_j} = \frac{df(tag_i \cap tag_j)}{min \quad df(tag_i), df(tag_j)} \tag{1.5}$$

- **Document frequency of the intersection of two tags over the maximum tag document frequency between them** [Equation (1.6)]: the weight is adjusted with the maximum tag document frequency of both tags in the dataset. In this weighting function we assume again the initial hypothesis, but unlike the preceding one, the values are scaled-down by the number of documents tagged with the most common tag.

$$W^{max}_{tag_i,tag_j} = \frac{df(tag_i \cap tag_j)}{max \quad df(tag_i), df(tag_j)} \tag{1.6}$$

Representation by Document Content

In order to represent a tag by content, we consider the documents that were annotated with that tag. Specifically, we limit to the textual content. Since each tag has many documents annotated, we merged the textual content of all those underlying documents. This approach was first introduced in Zubiaga et al. (2009).

However, we think we should not include all the tags in the same way in the document representation, as some of them may be hardly important because they have lower post count, and because of the associated computational cost. In order to decide which tags to consider relevant for a document, we need to set a threshold; in this manner, only tags with a higher post count than the threshold are selected. We consider the average post count (26) like our threshold, extrapolating the average in the collection to each and every single document (see Figure 2). Hence, working only with the top ranked tags could be more precise in order to discover document content semantics and to find relations among the tags in T140 set.

Figure 2. X axis represents the rank of a tag in the top list of tags of the annotated resources, whereas Y axis represents the average post count for each of the positions in the ranking. Note that the tag ranked first could be different from resource to resource. The dashed line means the average post count for every tag positions

from first to 25th. In consequence, only the tags with a higher post count than the average (above dashed line) were selected in representations by document content.

Then, each of the T140 tags is represented by its corresponding documents and instead of representing each and every document as a vector, we merge all the documents corresponding to a particular tag (hereafter super-documents). Thus, we obtain 140 super-documents representing the tags in T140. Since a document can belong to more than one super-document if it has been tagged with more than one of the 140 tags, then documents might represent more than one tag, and so we would be taking into account co-occurrence information in an implicit way, as we introduced above at the beginning of this section.

The next step is to represent each super-document into the vector space model. At this stage we follow the process described by Zubiaga et al. (2009) in order to represent tags using *TF-IDF* weighting function. The final result of this process are 140 term vectors, corresponding to each of the T140 tags, composed by 17,518 features each.

Tag Clustering

As a state-of-the-art clustering algorithm, we use Self-Organizing Maps (SOM) (Kohonen (1990, 2001)) to carry out the experimentation. SOM has proven to be an effective way not only to organize information, but also to visualize it, and even to allow content addressable searches (Rauber et al. (2002), Vesanto & Alhoniemi. (2000), Russell et al. (2002), Perelomov et al. (2002), Roh et al. (2003)).

Kohonen's Self-Organizing Maps are unsupervised neural network architectures that use competitive learning in order to produce a spatial-topological relationship between the reference vectors of each neuron in a Vector Space Model (VSM); after a training process, and depending on high dimensional input vectors. The neurons are arranged as a regular node grid, usually with 2 dimensions. Thus, after the training phase, similar inputs to the map will produce nearby outputs into the node grid.

The SOM size was set to 12x12, in order to obtain a square map with a number of neurons close to the number of tags (144 neurons, and 140 tags) with a rectangular lattice. In this way, we have at least one neuron per tag. We did not want to force tag grouping due to map size, that is, if the number of tags is greater than the number of neurons, then multiple tags must share the same neuron because there is no space enough to allocate them in separate ones.

During map training the initial learning rate was set to 0.1, the initial neighborhood was set to 12, equal to map width, and the number of training iterations was 50,000. These values were chosen measuring map quality with the Average Quantization Error (AQE) after several tests with different configurations. AQE measures the average distance between input vectors and their associated reference vectors in the map. Other issues about the SOM are the same as in the standard implementation SOMlib (Rauber et al., 2002).

Analysis

The different weighting functions based on co-occurrences produced very similar maps from the point of view of tag groups, but W function showed better grouping than the others from a qualitative point of view. As our main goal in this chapter is comparing approaches based on co-occurrences with a content based one, we chose the map generated using W function to represent the co-occurrences approach, simplifying the comparison process.

In order to analyze our results, content based map is shown in Figure 3, while co-occurrences based map is shown in Figure 4. In these maps each table cell is a SOM neuron, which may contain tags. Each tag is formatted in such a way that the bigger is the font, the higher the popularity of that tag in the dataset. As it was stated before, proximity on the map implies relatedness among the tags. Finally, throughout this section we will refer to neurons using these positions in the table, e.g. *(x/y)*, being *x* the row number and *y* the column number.

Before analyzing our results, it is worth to note that this dataset has a clear bias towards computer related topics. Because of this, we analyzed the maps in a deeper level of detail, assuming that some tags having different meanings are biased to their computer related one, e.g. tool & tools are related to programming and not to diy (*do it yourself*).

Therefore, analyzing and comparing the maps, several issues emerge:

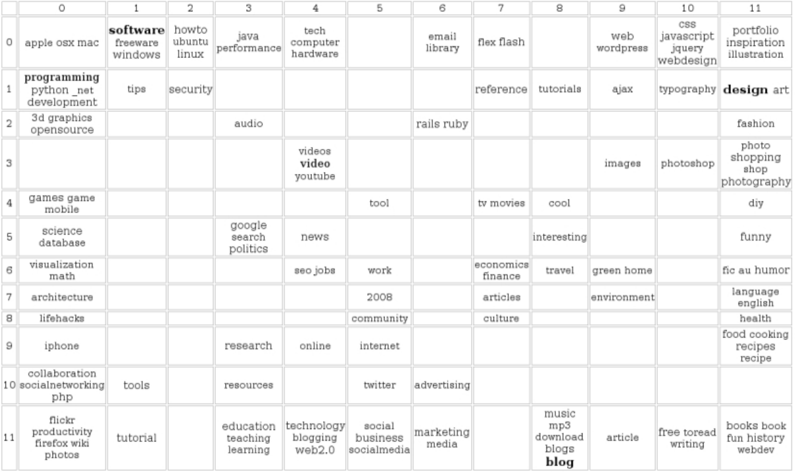
- Tags sharing the same stem not always share the same neuron in the content approach, e.g. blog (11/8) & blogging (11/4), fun (11/11) & funny (5/11), photo (3/11) & photos (11/0), tool (4/5) & tools (10/1), article (11/9) & articles (7/7). In the co-occurrences based map this fact only affects to one case, though they are in adjoining neurons: tool (1/1) & tools (1/2), appearing the others together.

- There are very related terms which appear together in the co-occurrences map, while they are separated in the content based, as: flickr (11/0), images (3/9), photo (3/11), photography (3/11) and photos (11/0); iphone (9/0) and mobile (4/0); google (5/3), search (5/3) and seo (6/4).

- In the content approach there are some strange groups, like: music, mp3, download, blog and blogs (11/8); book, books, fun, history and webdev (11/11). In the co-occurrences map some of these tags appear in groups that make more sense, e.g. css and webdev (6/0); audio, mp3 and music (11/8); fun, funny and humor (11/6).

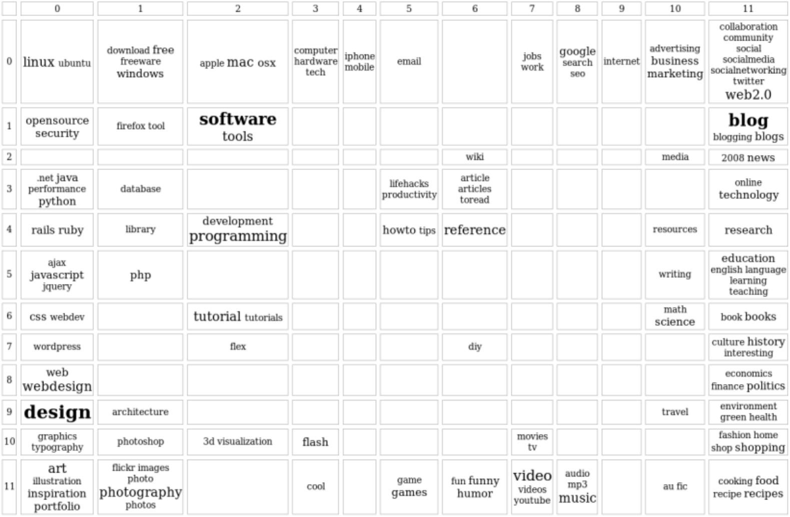
- Some tags which seem to be correctly grouped in the co-occurrences approach are located far away in the content side. This is the case of, for instance: fun (11/11), funny (5/11) and humor (6/11); blog (11/8), blogs (11/8) and blogging (11/4); webdev (11/11) and css (0/10); music (11/8), mp3 (11/8) and audio (2/3).

This analysis suggests that co-occurrences produce better groupings and therefore a better reorganized tag cloud. Our results show the high efficiency and accurate performance of approaches based on tag co-occurrences, which qualitatively outperform the grouping based on content. Furthermore, approaches based on co-occurrences greatly reduce the computational cost. Computing the tag co-occurrence values is feasible for tag clouds that contain about 100-200 tags. Accordingly, tag cloud reorganization is an affordable task that provides useful features to the browsing. This reorganization represents a good complement to traditional tag clouds, easing user navigation through big document collections. In consequence, relying on social data provided by end users shows to be a reliable source to find tag relations.



	0	1	2	3	4	5	6	7	8	9	10	11
0	apple osx mac	**software** freeware windows	howto ubuntu linux	java performance	tech computer hardware		email library	flex flash		web wordpress	css javascript jquery webdesign	portfolio inspiration illustration
1	**programming** python _net development	tips	security					reference	tutorials	ajax	typography	**design** art
2	3d graphics opensource			audio			rails ruby					fashion
3					videos **video** youtube					images	photoshop	photo shopping shop photography
4	games game mobile					tool		tv movies	cool			diy
5	science database			google search politics	news				interesting			funny
6	visualization math				seo jobs	work		economics finance	travel	green home		fic au humor
7	architecture					2008		articles		environment		language english
8	lifehacks					community		culture				health
9	iphone			research	online	internet						food cooking recipes recipe
10	collaboration socialnetworking php	tools		resources		twitter	advertising					
11	flickr productivity firefox wiki photos	tutorial		education teaching learning	technology blogging web2.0	social business socialmedia	marketing media		music mp3 download blogs **blog**	article	free toread writing	books book fun history webdev

Figure 3. Content based tag cloud reorganization.



	0	1	2	3	4	5	6	7	8	9	10	11
0	linux ubuntu	download free freeware windows	apple mac osx	computer hardware tech	iphone mobile	email		jobs work	google search seo	internet	advertising business marketing	collaboration community social socialmedia socialnetworking twitter web2.0
1	opensource security	firefox tool	**software** tools									**blog** blogging blogs
2							wiki				media	2008 news
3	.net java performance python	database				lifehacks productivity	article articles toread					online technology
4	rails ruby	library	development programming			howto tips	reference				resources	research
5	ajax javascript jquery	php									writing	education english language learning teaching
6	css webdev		tutorial tutorials								math science	book books
7	wordpress		flex				diy					culture history interesting
8	web webdesign											economics finance politics
9	**design**	architecture									travel	environment green health
10	graphics typography	photoshop	3d visualization	flash				movies tv				fashion home shop shopping
11	art illustration inspiration portfolio	flickr images photo photography photos		cool		game games	fun funny humor	video videos youtube	audio mp3 music		au fic	cooking food recipe recipes

Figure 4. Co-occurrences based tag cloud reorganization

Applications

A reorganized tag cloud presents different applications as compared to the original one:

- Feed subscription. Within the new cloud, users not only could subscribe to a unique tag, but also they can subscribe to a neuron or even a group of neurons, which contains a set of related tags.
- Finding collection-specific relations among tags allows to discover user communities, or even temporal trends; the new visualization improves the way in which users can explore the whole document collection.
- Analyzing the evolution of tag relations over time could show interesting characteristics of each tag, e.g., whether a tag is temporarily popular.

FUTURE RESEARCH DIRECTIONS

In this work we performed a qualitative evaluation of tag cloud reorganization based on two different approaches for tag representation. An interesting work to corroborate our findings would be to carry out a quantitative evaluation relying on external evaluation measures. In order to use this kind of measures, a gold standard would be needed. In this sense, another interesting work would be to develop a benchmark to be used as a gold standard for evaluation purposes, because as far as we know, there is not any available for the research community at the moment.

Besides, an analysis on tag evolution throughout time could be done based on the progressive map updates, e.g., a tag like *news* may vary its neighborhood due to the trends of the news in a specific period.

CONCLUSION

In this book chapter, we have motivated the need of a tag cloud reorganization process, which applied to the traditional tag clouds can help the user navigate through related content easily. We have also covered two different approaches to represent tags: one based on document content and another based on tag co-occurrences. In the case of co-occurrences, we used four different tag weighting functions in order to obtain a value representing the degree of co-occurrence between tag pairs, building a set of input vectors, one per tag, containing the similarity values between the vector tag and the rest of the T140 tags. These four functions were chosen in order to establish a baseline for tag co-occurrence representation facing the comparison with the content-based representation, which is based on the TF-IDF term weighting function. We have shown that representing tags by co-occurrences yields more accurate clusters than representing them by content.

Summarizing, we have shown that relying on social data provided by end users is a reliable source to find tag relations. These relations allow the composition of a reorganized tag cloud where each tag is surrounded by other related tags, enhancing users experience in social tagging systems.

REFERENCES

Begelman, G., Keller, P., & Smadja, F. (2006). Automated tag clustering: Improving search and exploration in the tag space. *Proceedings of WWW '06 Collaborative Web Tagging Workshop*. (pp. 15-33).

Brooks, C. H., & Montanez, N. (2006). Improved annotation of the blogosphere via autotagging and hierarchical clustering. *In WWW '06: Proceedings of the 15th international conference on World Wide Web*. (pp. 625-632). ACM.

Chen, Y.-X., Santaía, R., Butz, A., & Therón, R. (2009). Tagclusters: Semantic aggregation of collaborative tags beyond tagclouds. In: Butz, A., Fisher, B. D., Christie, M., Krüger, A., Olivier, P., Therón, R. (Eds.), *Smart Graphics*. Springer, 5531, 56–67.

Dattolo, A., Eynard, D., & Mazzola, L. (2011). An integrated approach to discover tag semantics. In *Proceedings of the 2011 ACM Symposium on Applied Computing. SAC '11*. (pp. 814-820). ACM.

Gabrielsson, S., & Gabrielsson, S. (2006). *The Use of Self-Organizing Maps in Recommender Systems*. Master's thesis, Department of Information Technology at the Division of Computer Systems, Uppsala University.

Garcia-Silva, A., Corcho, O., Alani, H., & Gomez-Perez, A. (2011). Review of the state of the art: Discovering and associating semantics to tags in folksonomies. *Knowledge Engineering Review* 26 (4).

Golder, S. A., & Huberman, B. A. (2006). The Structure of Collaborative Tagging Systems. *Journal of Information Science*, 32(2), 198-208.

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

Kohonen, T. (1990). The self-organizing map. Proceedings of the IEEE, 78 (9), 1464-1480.

Kohonen, T. (2001). *Self-Organizing Maps*. Berlin, Germany: Springer.

Lehwark, P., Risi, S., & Ultsch, A. (2007), Visualization and Clustering of Tagged Music Data. *Proceedings 31st Annual Conference of the German Classification Society*. (pp. 673-680). Springer.

Li, B., & Zhu, Q. (2008). The determination of semantic dimension in social tagging system based on som model. *Proceedings of the Second International Symposium on Intelligent Information Technology Application. IITA '08*. (pp. 909-913). IEEE.

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

Perelomov, I., Azcarraga, A. P., Tan, J., & Chua, T. S. (2002). Using structured self-organizing maps in news integration websites. *Proceedings of the 11th International World Wide Web Conference*.

Rauber, A. and Merkl, D. and Dittenbach, M. (2002). The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data. *IEEE Transactions on Neural Networks*, 13(6), 1331-1341.

Roh, T. H., Oh, K. J., & Han, I. (2003). The collaborative filtering recommendation based on SOM cluster-indexing CBR. *Expert Systems with Applications*, 25 (3), 413-423.

Russell, B., Yin, H., & Allinson, N. M. (2002). Document clustering using the 1 + 1 dimensional self-organising map. *Proceedings of the Intelligent Data Engineering and Automated Learning—IDEAL 2002*. (pp. 167-174). Springer.

Sbodio, M. L., & Simpson, E. (2009). Tag Clustering with Self Organizing Maps. *HP Labs Technical Reports*.

Sen, S., Lam, S. K., Rashid, A. M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F. M., & Riedl, J. (2006). Tagging, communities, vocabulary, evolution. *CSCW '06: Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*. (pp. 181-190). ACM.

Simpson, E. (2008). Clustering tags in enterprise and web folksonomies. *HP Labs Technical Reports*.

Smith, G. (2008). *Tagging: people-powered metadata for the social web*. Berkeley, California: New Riders.

Specia, L. & Motta, E. (2007). Integrating folksonomies with the semantic web. In Franconi, E., Kifer, M. & May, W. (Ed.), *The Semantic Web: Research and Applications* (pp. 624-639). Berlin: Springer.

Vandic, D., van Dam, J.-W., Hogenboom, F., & Frasincar, F. (2011). A semantic clustering-based approach for searching and browsing tag spaces. *Proceedings of the 2011 ACM Symposium on Applied Computing. SAC '11*. (pp. 1693-1699). ACM.

Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*. 11(3), 586-600.

Zubiaga, A., García-Plaza, A. P., Fresno, V., & Martínez, R. (2009). Content-based clustering for tag cloud visualization. *ASONAM '09: Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining*. (pp. 316-319). IEEE Computer Society.

ADDITIONAL READING SECTION

Babinec, M. S., & Mercer, H. (2009). *Metadata and open access repositories*. Philadelphia, PA: Taylor & Francis.

Baca, M., & Getty Research Institute. (2008). *Introduction to metadata*. Los Angeles, CA: Getty Research Institute.

Bonino, S. (2009). *Social tagging as a classification and search strategy: A smart way to label and find web resources*. Saarbrucken, Germany: VDM Verlag Dr. Mull.

Caplan, P. (2009). *Metadata fundamentals for all librarians*. New Delhi: Indiana Pub. House.

Foulonneau, M., & Riley, J. (2008). *Metadata for digital resources: Implementation, systems design and interoperability*. Chandos information professional series. Oxford: Chandos.

Gartner, R., L'Hours, H., & Young, G. (2008). *Metadata for digital libraries: State of the art and future directions*. Bristol: JISC.

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

Granitzer, M., Lux, M., & Spaniol, M. (2008). *Multimedia semantics: The role of metadata*. Berlin: Springer.

Heymann, P., Koutrika, G., & Garcia-Molina, H. (2008). Can social bookmarking improve web search?. *WSDM '08: Proceedings of the international conference on Web search and web data mining* (pp. 195-206). New York, NY, USA: ACM.

Hider, P. (2009). *Information resource description: Creating and managing metadata*. London: Facet.

Hillmann, D. I., Guenther, R., Hayes, A., Library of Congress., & Association for Library Collections & Technical Services. (2008). *Metadata standards & applications*. Washington, D.C: Library of Congress.

International Conference on Metadata and Semantics Research, Sartori, F., Sicilia, M.-A., & Manouselis, N. (2009). *Metadata and semantic research: Third international conference*, MTSR 2009, Milan, Italy, October 1-2, 2009 : proceedings. Berlin: Springer.

Lanius, L., & Vermont. (2009). *Embracing Metadata: Understanding MARC and Dublin Core [workshop]*. Montpelier, Vt: Vermont Dept. of Libraries.

Mittal, A. C. (2009). *Metadata management*. Delhi, India: Vista International Pub. House.

Noll, M. G., & Meinel, C. (2008a). Exploring social annotations for web document classification. *Proceedings of the 2008 ACM symposium on Applied computing* (pp. 2315-2320). Fortaleza, Ceara, Brazil: ACM.

Paukkeri, M., Pérez García-Plaza, A., Fresno, V., Martínez, R., & Honkela, T. (2012). Learning a taxonomy from a set of text documents. *Applied Soft Computing, 12*(3), 1138–1148.

Pérez García-Plaza, A., Fresno, V. & Martínez, R. (2008) Web Page Clustering Using a Fuzzy Logic Based Representation and Self-Organizing Maps. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*: Vol. 1 (pp. 851-854).
Sydney, Australia, Dec. 9-12, 2008.

Peters, I. (2009). *Folksonomies. Indexing and Retrieval in Web 2.0 (Knowledge & Information: Studies in Information Science)*. Berlin, Germany: De Gruyter/Saur.

Ramage, D., Heymann, P., Manning, C. D., & Garcia-Molina, H. (2009). Clustering the tagged web. *Proceedings of the Second ACM International Conference on Web Search and Data Mining* (pp. 54-63). Barcelona, Spain: ACM.

Smith, G. (2008). *Tagging: people-powered metadata for the social web*. Berkeley, California, United States: New Riders.

Taylor, A. G., & Joudrey, D. N. (2009). *The organization of information*. Westport, Conn: Libraries Unlimited.

Turrell, A. (2008). *Augmenting Classifications and Search with Tags to Create Usable Content- and Product-Based Websites*. Baltimore, MD: University of Baltimore.

Wu, X., Zhang, L., & Yu, Y. (2006). Exploring social annotations for the semantic web. *WWW '06: Proceedings of the 15th international conference on World Wide Web* (pp. 417-426). New York, NY, USA: ACM.

Zhou, D., Bian, J., Zheng, S., Zha, H., & Giles, C.L. (2008). Exploring social annotations for information retrieval. *Proceedings of the 17th international conference on World Wide Web* (pp. 715-724). Beijing, China: ACM.

Zubiaga, A., Martínez, R., & Fresno, V. (2009). Getting the most out of social annotations for web page classification. *DocEng'09: Proceedings of the 9th ACM symposium on Document engineering* (pp. 74-83). New York, NY, USA: ACM.

KEY TERMS & DEFINITIONS

- **Tagging:** Tagging is an open way to assign tags or keywords to resources or items (e.g., web pages, movies or books), in order to describe them. This enables the later retrieval of the resources in an easier way, using tags as resource metadata. As opposed to a classical taxonomy-based categorization system, they are usually non-hierarchical, and the vocabulary is open, so it tends to grow indefinitely. For instance, a user could tag this chapter as social-tagging, research and chapter, whereas another user could use web2.0, social-bookmarking and tagging tags to annotate it.

- **Social tagging:** A tagging system becomes social when its tag annotations are publicly visible, and so profitable for anyone. The fact of a tagging system being social implies that a user could take advantage of tags defined by others to retrieve a resource.
- **Social bookmarking:** Delicious, StumbleUpon and Diigo, amongst others, are known as social bookmarking sites. They provide a social means to save web pages (or other online resources like images or videos) as bookmarks, in order to retrieve them later on. In contrast to saving bookmarks in user's local browser, posting them to social bookmarking sites allows the community to discover others' links and, besides, to access the bookmarks from any computer to the user itself. In these systems, bookmarks represent references to web resources, and do not attach a copy of them, but just a link. Note that social bookmarking sites do not always rely on social tags to organize resources, e.g., Reddit is a social bookmarking approach to add comments on web pages instead of tags. The use of social tags in social bookmarking systems is a common approach, though.
- **Tag Cloud:** In order to enable visual browsing, social bookmarking systems typically provide an interface model known as tag cloud. These clouds are one of the main ways of browsing and discovering web documents on social bookmarking systems, as a structure that provides a visual summary of the most popular topics in the collection. Tag clouds comprise between 50 and 200 of the most popular tags on the site, where the more popular is a tag, the bigger it is shown. Sometimes, tags are sorted alphabetically, randomly, or using other non-semantic orderings.
- **Folksonomy:** As a result of a community tagging resources, the collection of tags defined by them creates a tag-based organization, so-called folksonomy. A folksonomy is also known as a community-based taxonomy, where the classification scheme is plain, there are no predefined tags, and therefore users can freely choose new words as tags. A folksonomy is basically known as weighted set of tags, and may refer to a whole collection/site, a resource or a user. A summary of a folksonomy is usually presented in the form of a tag cloud.
- **Simple tagging:** users describe their own resources or items, such as photos on Flickr, news on Digg or videos on Youtube, but nobody else tags another user's resources. Usually, the author of the resource is who tags it. This means no more than one user tags an item. In many cases, like in Flickr and Youtube, simple tagging systems include an attachment to the resource, and not just a reference to it.
- **Collaborative tagging:** many users tag the same item, and every person can tag it with their own tags in their own vocabulary. The collection of tags assigned by a single user creates a smaller folksonomy, also known as personomy. As a result, several users tend to post the same item. For instance, CiteULike, LibraryThing and Delicious are based on collaborative tagging, where each resource (papers, books and URLs, respectively) could be tagged (therefore annotated) by all the users who considered it interesting.