Content-based Clustering for Tag Cloud Visualization

Arkaitz Zubiaga, Alberto P. García-Plaza, Víctor Fresno, Raquel Martínez
Dpto. Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia
Madrid, Spain
{*azubiaga, alpgarcia, vfresno, raquel*}*@lsi.uned.es*

***Abstract*—Social tagging systems are becoming an interesting way to retrieve web information from previously annotated data. These sites present a tag cloud made up by the most popular tags, where neither tag grouping nor their corresponding content is considered. We present a methodology to obtain and visualize a cloud of related tags based on the use of self-organizing maps, and where the relations among tags are established taking into account the textual content of tagged documents. Each map unit can be represented by the most relevant terms of the tags it contains, so that it is possible to study and analyze the groups as well as to visualize and navigate through the relevant terms and tags.**

***Keywords*-social-tagging; clustering; information access; visualization;**

I. Introduction

In social bookmarking sites people can post and tag already posted content with their preferred tags, so it could be expected that the more users describe an item, the more representative is its tag set. In this context, several methods and approaches have been proposed to improve tasks such as: search and navigation strategies and results, tag cloud visualization, and recall and precision in feed subscription services, etc. All of them consider tag co-occurrence to organize related tags into clusters or groups, whereas some of them use extra information from the users, additional resources and the Semantic Web. As far as we know, there are no works using textual content of the annotated web documents to extract the relations among tags.

In this paper, we present a methodology to identify inter-related tags based on the textual content of tagged web documents by means of Self-Organizing Maps (SOM). It allows social tagging sites to suggest tags based on the neighborhood in the map, as well as to improve feed subscription services for related tag sets and the extraction of the most relevant terms for each group of tags by means of language modeling techniques. Therefore, the resulting SOM turns into a richer tag cloud that provides an alternative way to visualize and navigate through tags and terms.

This paper is organized as follows. Section II reviews different methods and approaches to identify inter-related tags. In Section III, we introduce the dataset generated for this work, explaining how our methodology works and the algorithms and techniques used. Section IV shows and analyzes experimental results after tag clustering. Finally, in Section V, the main conclusions of this approach are presented, and future work is proposed.

II. Related Work

Several methods and approaches have been proposed to identify inter-related tags, considering tag co-occurrence to organize related tags into clusters ([7], [2]). In [7], the author obtains a subsumption based model derived from the co-occurrence of tags to find groups of related tags from Flickr. In [2], they build an undirected graph representing the tag space, where the vertices correspond to tags, and edges between them represent their co-occurrence frequency. They obtain clusters of related tags, but since some clusters are too large, they apply a spectral clustering algorithm to refine them. In [10], use information from the co-occurrence of tags, resources and users in a probabilistic model to generate groups of semantically related tags. [5] uses a tripartite model involving users (*actors*), tags (*concepts*) and resources (*instances of concepts*) and builds graphs relating tags with both users and resources. Other works try to identify semantic relations using ontologies [1] and the Semantic Web. In [8] the authors derive meaningful groups of tags corresponding to concepts in ontologies by means of co-occurrence analysis and clustering techniques. The relations within tags in each cluster are discovered by combining the Semantic Web and resources such as Wikipedia or Google. Based on this approach, [1] only rely on online ontologies to obtain semantic enrichment of folksonomy tags.

In addition to these works, most of the tagging systems provide functionalities to show groups or clusters and relations among tags, which apparently rely on co-occurrence information and clustering techniques, but do not provide detailed information about the methodologies they use.

III. Our Methodology

Present work introduces a methodology to organize and visualize the tag cloud making it easier to analyze relations between tags and their content. Our methodology involves several steps: a) Compilation of a dataset and selection of relevant tags; b) Tag representation based on tagged documents content; c) Clustering with SOMs to organize

and visualize tags; and d) Extracting outstanding terms for every group using language modeling techniques.

A. The DeliciousT140 Dataset

We collected a dataset from Delicious during June 2008, called 'DeliciousT140', to carry out our experimentation and to validate our methodology. We took the 140 most popular tags from the site, that is, the whole tag set on its tag cloud, also referred to as T140. Based on this, we monitorized the latest posts for each tag in the T140 set, and we obtained $379,931$ unique documents. After that, we queried Delicious for getting the post count and weighted common tags list for each item. At the same time, we crawled all these items in order to get their content.

Finally, we filtered the collection by language, getting only the english-written documents. Additionally, since Delicious only shows as common tags those posted at least twice, we also reduced the collection to the documents with common tags on its set of tags, filtering the rest. This leaded $144,574$ documents, on which $67,104$ different tags had been set. The dataset is available as a benchmark[1].

B. Tag Representation

Each document (item) is composed by a set of weighted tags. These weights mean the number of users decided to assign the tags to the document and, as a consequence, their agreed relevance for the document. Hence, we should not include all of them in the document representation, because some of them may be hardly important due to their low post count. In order to decide which tags to consider relevant for a document, we set a threshold; only tags with a higher post count than the average were selected.

The top ranked tags are highly posted comparing to the rest. Hence, we believe that working only with the top ranked tags could be more precise in order to discover document content semantics and find relations among the T140 set. Tags above that value were considered representative for a document only if they were in the T140 set. Then, a document was not included in any tag when all its representative tags were not in T140. It is noteworthy that the resulting dataset is not balanced, e.g. the most represented tag is *design* with $11,856$ documents, whereas the tag with the least documents is *2008*, with 219 documents.

At this point, we got a filtered subcollection. Each of the T140 tags was assigned its corresponding documents. Instead of representing each and every document as a vector, we merged all the documents corresponding to a particular tag. Thus, we got 140 super-documents representing the 140 most common tags. Since a document can be included in more than one super-document (if it has been tagged with more than one of the T140 tags) we are taking into account co-occurrence information in a implicit way.

To represent each super-document into the vector space model, we first removed the HTML format to extract the plain text, we removed the most common stopwords, using an english stoplist, and accomplished a stemming phase with the Porter algorithm. After applying the *tf-idf* weighting, a dimensionality reduction stage was carried out to reduce vectors dimension: we removed the terms with highest and lowest document frecuency (*df*) values [3].

C. SOM-based Tag Clustering

After the representation process, we aimed to group tags sharing similar content. We decided to use SOMs [4] for the clustering process, since it has proven to be a good way to organize information and visualize it, and even allows content addressable searches [9]. Kohonen's SOMs are unsupervised neural networks. They produce a spatial-topologic relationship among the reference vectors of each neuron after a training process, and depending on input vectors (in our case, the 140 term vectors corresponding to the T140 set) with the same dimensionality than the reference vectors. The winner neuron is the node with smallest distance to the input sample. The aim is ordering these data according their mutual similarities making iterative comparisons with the input data collection.

In our experiments 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. In this way, we have at least one neuron per tag. During map's training the initial learning rate was set to 0.1, the initial neighborhood was set to 12, and the number of training iterations was 50000. These values were chosen measuring map's quality with the Average Quantization Error (AQE) after several tests with different configurations. Other issues about the SOM are the same as in the standard implementation SOMlib[2].

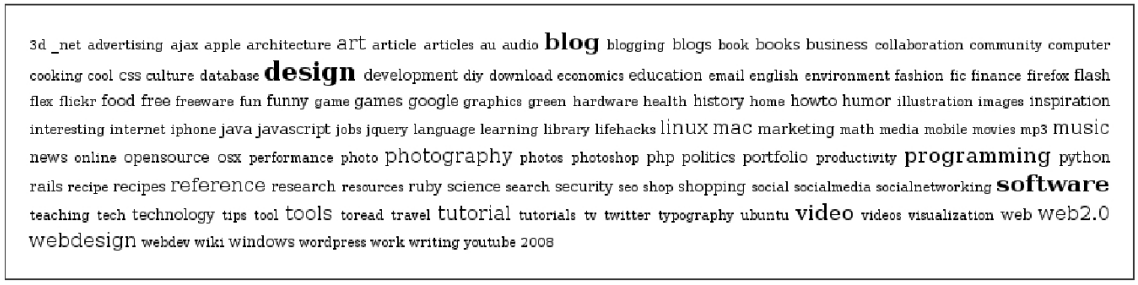
D. Terminology Extraction by Language Models

Once the map is calibrated and the tags have been grouped in neurons, we can extract the representative terminology for each of the neurons. This allows to extract terminology related to neurons with no more than a tag, or neurons that group several tags. First, we grouped all the documents corresponding to the tags labeling a neuron, and we generated a ranked list of terms. To accomplish this task, we decided to use language modeling techniques, applying the KLD weighting function to determine terms in a neuron diverging to the rest of neurons in the map:

KLD $= P_n(t).Log\frac{P_n(t)}{P_m(t)}$; where P_n is the occurrence probability of the term t in the neuron, and $P_m(t)$ is the occurrence probability of t in the whole map. A larger KLD score for a term means that the term is more discriminative.

[1] http://nlp.uned.es/social-tagging/delicioust140/

[2] http://www.ifs.tuwien.ac.at/∼andi/somlib/



	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 1. Original tag cloud (above) and tag cloud clustered by content (below)

IV. Experiments and results

As a result of the experiments we obtained the map shown in Figure 1.

The labeled map allows us to visualize tag distribution. Nearby areas in the map means similarity among tags content. The fact that a neuron groups several tags means that their contents are similar. Anyway, neighbors in the map are not equidistant. Our map is a reorganization of the original Delicious tag cloud, providing a new cloud grouped by tag content and a term list to allow deeper navigation possibility. We obtained a list for each neuron, and a second list for each tag. Due to the lack of space we cannot show an example here[3].

Some ideas can be extracted analyzing tag distribution. Some neurons show obvious tag relations, such as neuron (7/0) with *flex* and *flash* tags; neuron (6/2) with *rails* and *ruby* tags; neuron (4/3) with *videos*, *video* and *youtube* tags; neuron (11/9) with *food*, *cooking*, *recipe* and *recipes* tags. There are some neurons where users with a deep knowledge on the domain would deduce semantic relations; neuron (2/0) contains the tags: *howto*, *ubuntu* and *linux*.

Some unobvious relations can also be found. An example of this type of relations is neuron (0/4), containing *games*, *game* and *mobile* tags, which discovers mobile games as a subject of interest for a specific community. In the same way, neuron (0/5) relates *science* and *database* tags. After analyzing the relevant terms we discovered there is a community interested in scientific databases. Finally, it is worth to comment neuron (3/5), with *google*, *search* and *politics*, probably due to the time interval when the collection was retrieved. This neuron shows terminology related to US elections.

[3]Full map is available in http://nlp.uned.es/social-tagging/asonam2009/

For the T140 tags, a user could categorize most of them as belonging to different general categories. The categories we have found include: **Computer Science**, represented in grouped neurons (0/0), (0/1), (0/2), (1/0), (1/1), (2/0), (2/1), (3/0) and (4/0); **Graphical & Web Design**, represented in grouped neurons (9/0), (9/1), (10/0), (10/1), (11/0), (11/1), (11/2), (11/3), (10/3) and (9/3); **Education**, represented in grouped neurons (3/9), (3/10) and (3/11); **Cooking**, neurons (11/8) and (11/9); **Entertainment**, neurons (9/11), (10/11), (11/11), (11/5) and (11/6); and **Economics**, being neuron (7/6) and the grouped neurons (5/11) and (6/11). So, in these cases, the content similarity is related to the semantic similarity among tags.

Nevertheless, some tags seem to be incorrectly grouped; this is the case of the neuron (11/11), grouped in the entertainment topic, where *webdev* is grouped with *book*, *books*, *fun* and *history*. Neuron (0/11) also contains heterogenous thematic tags such as *productivity*, *firefox*, *wiki*, *photos* and *flickr*. Finally, we can find two topical groups in other neurons, such as (8/11) with *music*, *mp3*, *download*, *blog* and *blogs* tags. Probably blog is a very generic tag and could be easily combined with other tags, sharing most of the documents and, therefore, several terms with those tags.

Others are properly grouped in neurons, but without evident relation with their neighbors. Neuron (4/3) with *video*, *videos*, and *youtube*, is next to neuron (5/4) with *tool*. It could be a dimensionality-related consequence; there may not be space enough to separate them. Taking into account the type of tags, we can observe that many of the non-topical tags (generic, subjective), are alone in their neuron, such as: (8/4) *cool*, (8/5) *interesting*, (6/10) *advertising*, (1/11)*tutorial*, (4/9) *online*, (5/7) *2008*, (5/4) *tool*.

Finally, there are some cases where a tag in the singular is grouped far away from the same tag in the plural, e.g, tutorial (1/11) and tutorials (8/1). When there are documents tagged by one of them, but they are not tagged by both, then these documents will make the difference.

V. Conclusions and Outlook

In this paper we have presented a methodology to obtain and visualize a cloud of grouped tags based on the use of SOMs, and language models. Moreover, we have compiled a dataset based on Delicious as a benchmark. Our approach allows us to ease the discovery of relations between tags considering documents content and improves tag navigation by means of clustered tags and their relevant terminology. Our methodology facilitates the analysis and identification of tagging trends, interest communities, and semantic relations between tags by means of the study of the resulting map and the associated relevant terms.

An interesting application of our methodology would be to add a new functionality to the traditional tag cloud integrating it in a real system such as Delicious. It would allow a user to subscribe to a desired term in a tag. A user would not like to receive all the pages tagged with a particular tag, but only those containing a concrete term. In this case, the system could inform user even of the documents containing the selected term mapped in the same neuron as the tag. On the other hand, an analysis on tag evolution over time could be done based on the progressive map updates, e.g. a tag like "news" may vary its neighborhood due to the tendency of the news in a specific period, such as US elections.

Future work will include: (1) to perform a deeper study on the semantic nature of each tag; (2) to compare our results with an approach based on tag co-occurrence; (3) to apply a clustering algorithm to the SOM in order to obtain clusters; (4) to carry out a deeper analysis of the factors that influence the grouping (document representation, SOM parameters, etc.) to improve the resultant map; and (5) to apply this methodology to multilingual web resources.

Acknowledgment

This work has been supported by the Research Network MAVIR(S-0505/TIC-0267), and by the Spanish Ministry of Science and Innovation project QEAVis-Catiex(TIN2007-67581-C02-01).

References

[1] S. Angeletou, M. Sabou, L. Specia, and E. Motta. Bridging the gap between folksonomies and the semantic web. In *Proceedings of ESWC 2007: Workshop on Bridging the Gap between Semantic Web and Web 2.0*, 2007.

[2] G. Begelman, P. Keller, and F. Smadja. Automated tag clustering: Improving search and exploration in the tag space. In *WWW '06: Collaborative Web Tagging Workshop*, 2006.

[3] M. Dittenbach, D. Merkl, and A. Rauber. The growing hierarchical self-organizing map. pages 15–19. IEEE Computer Society, 2000.

[4] T. Kohonen. *Self-Organizing Maps*. Self-organizing maps. 3rd ed. Berlin: Springer, 2001, xx, 501 p. Springer series in information sciences, ISBN 3540679219, 2001.

[5] P. Mika. 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, 2007.

[6] B. Russell, H. Yin, and N. M. Allinson. Document clustering using the 1 + 1 dimensional self-organising map. In *Proceedings of the Third IDEAL*, 2002.

[7] P. Schmitz. Inducing ontology from flickr tags. In *WWW '06: Collaborative Web Tagging Workshop*, 2006.

[8] L. Specia and E. Motta. Integrating folksonomies with the semantic web. *The Semantic Web: Research and Applications*, 2007.

[9] J. Vesanto and E. Alhoniemi. Clustering of the self-organizing map. *IEEE-NN*, 2000.

[10] X. Wu, L. Zhang, and Y. Yu. Exploring social annotations for the semantic web. In *Proceedings of WWW '06*, pages 417–426, New York, NY, USA, 2006. ACM.