Investigating Community Structure In Social Tagging Systems

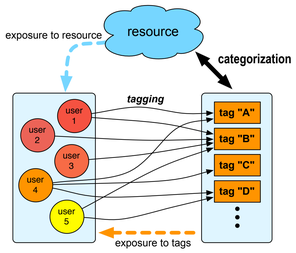
Ciro Cattuto[1,2], *Andrea Baldassarri*[2], *Vito D. P. Servedio*[2,1] *and Vittorio Loreto*[2]

[1] Centro Studi e Ricerche "Enrico Fermi"
Compendio Viminale, 00184 Roma, Italy

[2] Dipartimento di Fisica, Università di Roma "La Sapienza"
p.le Aldo Moro 2, 00185 Roma, Italy

Introduction

Collaborative tagging systems have been quickly gaining ground on the World Wide Web. In web-based applications like *del.icio.us*, *Flickr*, *CiteULike*, *BibSonomy* users enrich diverse resources, ranging from photographs to scientific references and web pages, with semantically meaningful information in the form of text labels, or "tags". Tags are freely chosen and users associate tags with resources in a totally uncoordinated fashion, for their own use. The tagging activity of each user is globally visible to the user community and the tagging process develops genuine social aspects and complex interactions. Despite the selfish nature of users' behavior, tagging systems exhibit cooperative dynamics leading to a bottom-up categorization ("folksonomy") of resources, shared throughout the user community [1,2].



Our analysis focuses on ***del.icio.us*** (http://del.icio.us/) for several reasons:

- it was the first system to deploy the ideas of collaborative tagging and it has acquired a paradigmatic character, making it a natural starting point for quantitative studies.
- it has been enjoying a large popularity, it has a large community of active users and it comprises a precious body of raw data on the static and dynamical properties of a folksonomy.
- it is a "broad folksonomy", i.e. single tagging events (posts) retain their identity and can be individually retrieved.

Experimental Data

We consider a set of resources and check whether the uncoordinated tagging activity of users is able to structure such a resource space in a semantically meaningful way, and whether such structures are accessible by using unsupervised methods. In order to perform this experiment we build a set of resources by merging two sets of 200 resources each: the first set contains resources tagged with "design" while the second set contains resources tagged with "politics". In this way we construct a dataset with at least two well-defined semantic regions.

For each resource in the dataset, we retrieve the entire sequence of user annotations (posts), i.e. the tags associated by each user with the resource. The tag vocabulary for the chosen set of 400 resources is shown below by means of a representation knows as "tag cloud": the size of each tag is proportional to the logarithm of the frequency of occurrence of that tag in the dataset.

37signals accessibility activism agency ajax animation apple art artist awards blog blogger blogging blogs book books browser bush business code color colors colour community computer cool creative creativity css culture daily design desktop desktops development diseño download ebooks economics education election entertainment extension extensions finance firefox flash flickr font fonts free freelance freeware fun funny gallery game games generator google government graphic graphics howto html humor icons ideas illustration image images imported inspiration interesting interface internet javascript law layout library lightbox links literature logo mac magazine maps mashup media money movies music news online opensource osx paper patterns photo photography photos photoshop php politics portal portfolio pricing productivity programming reading reference research resource resources search searchengine shopping showcase software stock system:unfiled tech technology template templates testing theme themes tips tool tools toread tutorial tutorials typography ui usability useful utilities video wallpaper wallpapers web web2.0 webdesign webdev website wordpress work xhtml

Sixth Framework Program
Information Society Technologies
Future and Emerging Technologies
contract number 34721
http://www.tagora-project.eu/

Distance Metrics

Users associate tags with resources. For every resource in the dataset, we retrieve all tags and measure their frequencies in the context of a resource. The tagclouds computed for a pair of sample resources is shown above. The size of a tag is proportional to the logarithm of its frequency. Red tags are shared by the two resources.

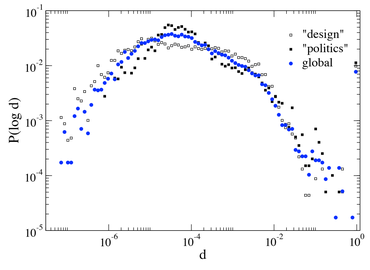
We introduce a weighted network of resources, and define the weight of an edge linking two resources R_1 and R_2 as:

$$w_{R_1,R_2} = \frac{\sum_{t\in T_1\cap T_2} \frac{min(f_t^1, f_t^2)}{f_t}}{\sum_{t\in T_1\cap T_2} \frac{max(f_t^1, f_t^2)}{f_t} + \sum_{t\in T_1 - T_2} \frac{f_t^1}{f_t} + \sum_{t\in T_2 - T_1} \frac{f_t^2}{f_t}}$$

f_t^1 frequency of tag *t* in the context of resource R_1

f_t^2 frequency of tag *t* in the context of resource R_2

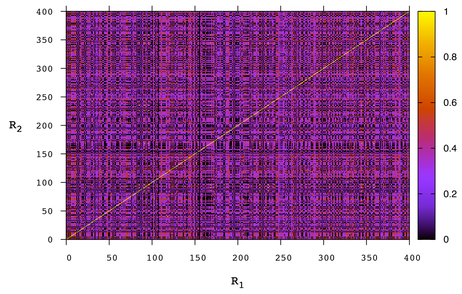
f_t global frequency of tag *t* , for all resources in the dataset



The figure above shows the strength distribution of resource pairs for three different sets of resources: the set of resources tagged with "design", the set of resources tagged with "politics" and their union. Because of the high dynamic range in link strengths, we enhance the contrast of our weight metric by raising it to small power. This is qualitatively equivalent to taking the logarithm of the weight, but is well-behaved in the vicinity of zero.

$$w'_{R_1,R_2} = (w_{R_1,R_2})^{0.1}$$

Spectral Analysis

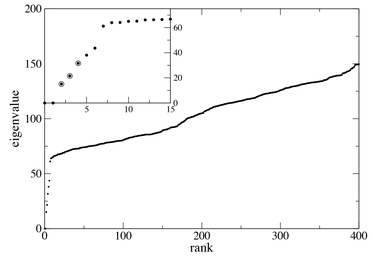


Weighted adjacency matrix w' for the full set of 400 resources. Each entry in the matrix corresponds to the strength w' of the link between two resources. The resources are randomly ordered and no structures are visible in this representation.

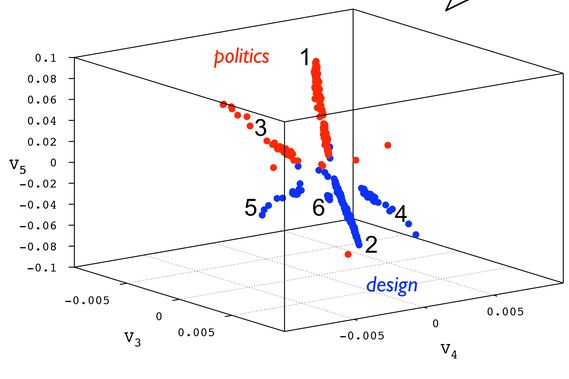
Detecting communities is equivalent to finding a permutation of the rows and columns of the weighted adjacency matrix *w'* that results in a clearly visible block structure along the main diagonal. We construct an auxiliary matrix *Q* and use information from its spectral properties to rearrange row and columns of the original matrix [3,4]. The matrix *Q* is non-negative and resembles the Laplacian matrix of graph theory.

$$W_{ij} = (1 - \delta_{ij})\, w'_{ij}$$

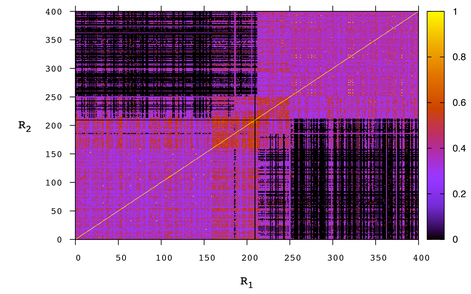
$$Q_{ij} = \left(\sum_j W_{ij}\right) - W_{ij}$$



We consider the three lowest non-trivial eigenvalues of the matrix *Q* and their associated eigenvectors (inset, above). On plotting the eigenvectors' components against each other in a three-dimensional space, resource communities emerge as well defined clusters of points (below).



Results



Once we have identified the communities, we permute the indexes of the original matrix *w'* so that the components of the same community are contiguous. The reordered matrix displays blocks along the diagonal, corresponding to resource communities.

Are communities we have found through the diagonalization of the matrix *Q* representing semantically separated areas in the space of resources? In order to check this we compute the distribution of tags over the resources belonging to each cluster, and below we show the corresponding tagclouds, in decreasing order of cluster size.

1
activism art blog burn bush creativity culture dvd economics flash freeware fun funny government history humor maps media money politics reference research software speechwriter statistics system:unfiled tools usa war windows

2
37signals art blog books css design development font fonts free graphics howto illustration inspiration photo photography photoshop portfolio productivity programming reference software system:unfiled themes tutorial tutorials typography web webdesign wordpress

3
activism blog blogs bush colbert comedy conservative culture election fraud freedom funny government grillo humor internet law libertarian maps media news political politics progressive science security system:unfiled usa video voting

4
art business color css design development flash free fun game games google graphics html inspiration patterns photography photos pricing reference resources search software stock system:unfiled tools web web2.0 webdesign webdev

5
ajax art awards blog blogger blogs color cool css design flash gallery graphics html images inspiration internet javascript lightbox politics portal portfolio reference system:unfiled templates tools web web2.0 webdesign webdev

6
ajax art blog books color css design desktop desktops development extension extensions firefox flash graphics icons illustration inspiration programming reference software system:unfiled technology tools typography wallpaper wallpapers web webdesign webdev

Conclusions

We analyzed the structure of the semantic space defined by a given subset of resources associated to two semantically different tags. We defined a suitable measure of similarity between resources, and used it to unravel the community structure of this semantic space. The two main resource communities, known in advance, were recovered entirely, together with non-trivial communities of smaller size. The presence of clearly detectable communities points in the direction of an effective cooperation among the users to efficiently map the semantic space with meaningul set of tags.

References

[1] T. Hammond, T. Hannay, B. Lund and J. Scott, *Social Bookmarking Tools (I): A General Review*, D-Lib Magazine **11**(4), (2005).

[2] S. Golder and B. A. Huberman, *Usage patterns of collaborative tagging systems*, Journal of Information Science **32**, 198 (2006).

[3] A. Capocci, V. D. P. Servedio, G. Caldarelli and F. Colaiori, *Detecting communities in large networks*, Physica A **352**, 669 (2005).

[4] M. E. J. Newman, *Finding community structure in networks using the eigenvectors of matrices*, Phys. Rev. E **74**, 036104 (2006).