Collaborative Tagging as a Tripartite Network

Renaud Lambiotte and Marcel Ausloos

SUPRATECS, Université de Liège, B5 Sart-Tilman, B-4000 Liège, Belgium
{Renaud.Lambiotte, Marcel.Ausloos}@ulg.ac.be

Abstract. We describe online collaborative communities by tripartite networks, the nodes being persons, items and tags. We introduce projection methods in order to uncover the structures of the networks, i.e. communities of users, genre families... The structuring of the network is visualised by using a tree representation. The notion of diversity in the system is also discussed.

1 Introduction

Recently, new kinds of websites have been dedicated to the sharing of people's habits and tastes, examples including their preferences in music, scientific articles, movies, websites... These sites allow members to upload from their own computer a library that characterises their habits in the corresponding topic (an iTunes music library for instance), and next to create a web page containing this list of items. Additionally, the website proposes the users to discover new content by comparing their taste with that of other users, thereby helping them discover new musics/books/websites... that should (statistically) fit their profile.

This method rests on a feedback between the users and a central server, and is usually called collaborative filtering. The emergence of these collaborative websites answers the needs of Internet users to retrieve useful and coherent informations from the millions of pages and data that form the Web. The main particularities of collaborative systems are: (i) their non-commercial purpose, even though the frontier with commercial companies is more and more vague (see for instance the acquisition of *del.icio.us* by *Yahoo* in November 2005); (ii) their transparency, namely these sites are relatively open and do not hide the profiles of each user, contrary to Amazon for instance. From a scientific point of view, this transparency opens perspectives in order to perform large scale experiences (including thousands of people) on taste formation, quantitative sociology, musicology... The available data also suggest alternative methods in order to perform large scale classifications of music/science/internet. Those subdivisions should be based on the intrinsic structure of the audience of the items.

In parallel with this sharing and statistical comparing of content, collaborative websites usually propose tagging possibilities. This process, called "folksonomy" (short for "folk taxonomy") means that the websites allow users to publicly tag their shared content, the key point being that their tag is not only accessible to themselves, but also to the whole ensemble of users. For instance, in the case of music sharing habits, a group like *The Beatles* is described in different ways,

V.N. Alexandrov et al. (Eds.): ICCS 2006, Part III, LNCS 3993, pp. 1114–1117, 2006.

i.e. *pop, 60s, britpop...*, that depend on the different backgrounds, tastes, music knowledge or *network of acquaintances...* of the users.

2 Methodology

The structure of collaborative websites can be viewed as a tripartite network. Namely, it is a network composed of three kinds of nodes: i) the persons or users μ; ii) the items i that can be music groups or scientific articles; iii) the tags I that are used by the person μ to describe the item i. Depending on the systems under consideration, a person can use one or several tags on each item. The resulting network can be represented by a graph where edges run between the item i and the user μ, passing through the tag I. Moreover, a weight is attributed to each link depending on the number of tags given by μ to i. For instance, if μ uses two tags for i, the weight of the links is $\frac{1}{2}$.

Let us note n_U the number of users, n_{It} the number of items, and n_T the number of tags in the considered sample. Consequently, each listener μ can be characterised by the $n_{It} \times n_T$ matrix $\overline{\overline{\sigma}}^{\mu}$:

$$\overline{\overline{\sigma}}^{\mu} = \begin{pmatrix} 0 & ... & 1/2 & ... & 1/2 & ... & 0 \\ ... & ... & ... & ... & ... & ... & ... \\ ... & 1/3 & ... & 1/3 & ... & ... & 1/3 \\ ... & ... & ... & ... & ... & ... & ... \end{pmatrix} \tag{1}$$

where $\overline{\overline{\sigma}}^{\mu}_{iI}$ denotes the weight of tag I in its description of i, so that $\sum_I \overline{\overline{\sigma}}^{\mu}_{iI} = 1$ if μ owns i and zero otherwise. Each item and each tag is also characterised by similar matrices that we note $\overline{\overline{\gamma}}^{i}$ and $\overline{\overline{\alpha}}^{I}$ respectively.

A common way to simplify the analysis of multi-partite networks consists in projecting them on lower order networks, i.e. unipartite or bipartite networks. In the following, we only focus on the correlations between two kinds of nodes, for instance between the users and the items. To do so, we first reduce the tripartite network to a bipartite one by summing over all nodes of one kind, thereby neglecting possible correlations between the three kinds of nodes. For instance, the bipartite network users-item is obtained by summing over all tags, so that each listener μ is now described by the the n_{It}-vector $\overline{\sigma}^{\mu}_{|_I}$:

$$\overline{\sigma}^{\mu}_{|_I} = (..., 1, ..., 0, ..., 1, ...), \tag{2}$$

the index running over all items, and where $\overline{\sigma}^{\mu}_{|_I} = \sum_I \overline{\overline{\sigma}}^{\mu}_{iI}$. The items are characterised by the n_U-vector $\overline{\gamma}^{i}_{|_I} = (..., 1, ..., 0, ..., 1, ...)$. These vectors are signatures of the users/items, that account for their interests/audience. In the case of music, we call these vectors the *music signatures* of people and groups.

In order to project the bipartite network on a unipartite one, we look at the correlations between two nodes of the same kind, relatively to his behaviour with another kind. For instance, one may look how persons μ and λ are correlated by using common items. To do so, we introduce the symmetric correlation measure:

$$C_{CF}^{\mu\lambda} = \frac{\overline{\sigma}^{\mu}_{|_I}.\overline{\sigma}^{\lambda}_{|_I}}{|\overline{\sigma}^{\mu}_{|_I}||\overline{\sigma}^{\lambda}_{|_I}|} \equiv \cos\theta_{\mu\lambda} \tag{3}$$

where $\overline{\sigma}^{\mu}_{|_I}.\overline{\sigma}^{\lambda}_{|_I}$ denotes the scalar product between the two n_{It}-vector, and $||$ its associated norm. This correlation measure, that corresponds to the cosine of the two vectors in the n_{It}-dimensional space, vanishes when the persons have no common item, and is equal to 1 when their item libraries are strictly identical.

At this level, the search for structures requires the analysis of large correlation matrices, and the uncovering of connected blocks that could be identified as families/genres/communities. In order to extract families of alike elements from the correlation matrix $\mathbf{C}$, we define the filter coefficient $\phi \in [0,1[$ and filter the matrix elements so that $C^{ij}_{\phi} = 1$ if $C^{ij} > \phi$ and $C^{ij}_{\phi} = 0$ otherwise. Starting from $\phi = 0.0$, namely a fully connected network, increasing values of the filtering coefficient remove less correlated links and lead to the shaping of well-defined islands, completely disconnected from the main island.

A branching representation of the community structuring [1] is used to visualise the process. To do so, we start the procedure with the lowest value of $\phi = 0.0$, and we represent each isolated island by a square whose surface is proportional to its number of internal elements. Then, we increase slightly the value of ϕ, e.g. by 0.05, and we repeat the procedure. From one step to the next step, we draw a bond between emerging sub-islands and their parent island. The filter is increased until all bonds between nodes are eroded (that is, there is only one node left in each island). Let us note that islands composed of only one element are not depicted for the sake of clarity. Applied to the above correlation matrix C^{ij}, the tree structure gives some insight into the specialisation by following branches from their source toward their extremity.

By construction, the above procedure unambiguously attributes to each element a hierarchical set of categories [2]. Consequently, starting from collaborative filtering that is a non-exclusive and non-hierarchical process, we have arrived to an exclusive and hierarchical structure that may be viewed as a taxonomy. This relation could have helpful applications in order to automatically structure content in systems without a central authority.

3 Applications: Measuring Diversity

Amongst others, this work provides tools in order to compare the tastes and interests of different persons, as well as to measure their *diversity*. Practically, let us focus on music collaborative websites and consider the case of two users μ_1 and μ_2 who own a list of music groups, each of them characterised by a spectrum of genres. From this knowledge, one would like to find a quantitative measure of the diversity of the persons, and a way to measure whether they have a similar taste. Let us note $\boldsymbol{\tau}_{\mu_1}$ and $\boldsymbol{\tau}_{\mu_2}$ the vector of genres characterising μ_1 and μ_2, where

$$\boldsymbol{\tau}_{\mu_1} = (\tau_{\mu_1 1}, ..., \tau_{\mu_1 I}, ..., \tau_{\mu_1 n_T}). \tag{4}$$

$\tau_{\mu_1 I}$ is the number of times that the tag I is associated to an item of μ_1 and n_T is the total number of tags in the system. A naive way to study diversity consists in implicitly assuming that all tags have different meaning and in characterising a person by the width of the distribution of τ. This is what we have done in ref.[3], where we defined a probabilistic entropy in order to measure these fluctuations. It is nonetheless an oversimplification that does not take into account the correlations between the tags, i.e. the fact that tags may have more or less equivalent meanings.

A more refine measure of diversity should require a proper counting of the *categories* to which the user belongs. To do so, we propose to visualise the branches and sub-branches of the hierarchical tree in which the user is more active than the average. Let us assume that, at some level of the filtering, an island (the node of one branch in the tree representation) is composed of K tags, say $I_1, ..., I_i, ..., I_K$. Let us denote $\tau^S_{I_i}$ the total number of times the tag I_i is used in the sample, while, as defined above, $\tau_{\mu I_i}$ is the total number of times I_i is tagged to the items belonging to μ. The above island, composed of K genres, is then characterised by:

- $p^S = (\sum_{i=1}^{K} \tau^S_{I_i})/(\sum_{I=1}^{n_T} \tau^S_I)$, that gives the empirical probability that a tag used in the sample belongs to the considered island.

- $p^\mu = (\sum_{i=1}^{K} \tau^\mu_{I_i})/(\sum_{I=1}^{n_T} \tau^\mu_I)$, that is the probability that a tag used on an item of μ belongs to the same island.

The activity of the user in the island is simply evaluated by looking at the ratio $r = p^\mu/p^S$. By construction, this quantity is bigger than 1 if the user owns many groups belonging to this island, and smaller than 1 otherwise.

Applying the method to all the nodes of the tree representation , and using a colour representation in order to represent the value of r, i.e. the nodes are printed in a colour ranging from green (low r) to blue (high r) [4], the user's diversity may be visualized. Moreover, different users may be compared by looking whether they are active in the same branches or in different branches.

During the poster presentation, the above techniques will be applied to empirical data extracted from websites specialised in music, e.g. *audioscrobbler.com* and *musicmobs.com*, and in scientific articles, i.e. *citeulike.com.*

Acknowledgments. This work has been supported by European Commission Project CREEN FP6-2003-NEST-Path-012864.

References

1. R. Lambiotte and M. Ausloos, *Phys. Rev. E*, **72**, 066107 (2005)
2. S. Golder and B. A. Huberman, cs.DL/0508082
3. R. Lambiotte and M. Ausloos, *EPJB*, in press; physics/0509134
4. R. Lambiotte and M. Ausloos, cs.DS/0512090