Blogosphere community formation, structure and visualization

Juan J. Merelo, Beatriz Prieto
Depto. Arquitectura y Tecnología de Computadores, U. Granada (Spain)
Fernando Tricas
Depto de Informática, Universidad de Zaragoza (Spain)

Abstract

Even as social networks have become fashionable over the last few years, the emphasis has been on "artificial" social networks (like Orkut, where you state your links explicitly) over "natural" social networks (where social links are deducted from user actions, i.e., those underlying the blogosphere, globally or regionally). Blog readers and writers form communities, and there are several tools that allow to visualize them. By mapping a subset of the blogosphere at different intervals in time, a picture of its evolution can be also drawn. And, finally, by looking closely at each community and its evolution, some conclusions can be drawn on what is the feature that defines such community. In this presentation, we will show which tools are available to find and map weblog communities, the evolution of how selected communities, and what conclusions we draw from it. In particular, we will present a neural-network based tool called Kohonen's map, which we have introduced for mapping and representing weblog communities.

Introduction and state of the art

Weblogs, or blogs, can be considered on-screen renderings of communities of readers/writers, which establish long-running relationships; these communities include weblog owners/writers or editors, people that post comments to weblog stories, and *silent* but persistent readers, both of whom might have their own weblog. A weblog by itself need not be important, but as part of a community, its importance cannot be disregarded. All weblogs in the world can be seen as components of a set of communities, each one with its own idols, axioms, enemies, and hierarchies. Communities are not clear-cut, since a particular weblog might belong to several communities at the same time, even though most weblogs (in fact, all weblogs in the Spanish-speaking community (Tricas et al., 2003) are connected to each other by a finite set of links.

Since blogs perform a sort of collaborative filtering of information published on the web at large, and are starting to be used as knowledge management tools, identifying communities becomes specially important. Information flows more easily within communities than outside them; getting a message across to as many persons as possible becomes, then, a matter of identifying communities, and the position of different sites within them. As straightforward as this view of the community concept might seem, the main problem is that there is no universally accepted definition of community in complex networks. Informally, it can be defined as a set of blogs (or websites) that share common interests, but this only begs the definition of *common* and *interest*. Another possible definition is to consider a community as a set of blogs that have a stronger relationship among them than with the rest of the websites of the same class. Equating *relationship* with *hyperlinks* means that a community is

a set of weblogs that has more links within the group than to outside sites. However, while heavily linking implies belonging to the same community, the inverse does not necessarily hold: two weblogs (and its readers/commenters; from now on, every time we refer to weblogs in a community context, we actually refer to the group of persons related to that weblog: readers, writer(s), commenters, and even those who link to it without even reading it) might both link to the same one, and thus belong, in a sense, to the same community without being aware of each other or the community.

In practice, data available to discover community ascription must be included in the web page source code, which is text formatted using HTML tags and some additional meta-tags; sometimes, each text can be assigned a time-stamp. The aforementioned common interest will have to be identified by using this data. From the point of view of text content, two websites are related if they deal approximately with the same topics. Considering links, two websites are related if they link to each other in either direction. These two definitions are actually correlated: Menczer has proved that pages that link to each other are semantically related. Furthermore, there are several additional problems with communities related by content: if a community is defined by keywords, synonyms and hypernims, if not considered or appropriately chosen, can lead to overseeing certain websites. This problem is aggravated further by the distinct characteristics of weblogs as rapidly changing websites and not focusing on a single topic or set of topics. Using content requires a vector space representation, usually term frequency/inverse document frequency. This representation is usually highly-dimensional, much more so than using links to other members of the set of webs that is going to be studied. For a small set of sites, link-based representation is much more compact. Relationship expressed by content distance, however, is implicit: two weblogs talking about politics, for instance, need not know each other, although it is very likely that they do since at least the Spanish blogosphere is connected (Tricas et al., 2003). Moreover, in many cases, communities are multilingual; two weblogs closely related to each other (for instance, written by the same author) but written in different languages (for instance, Spanish and Catalan, or Spanish and English) will be completely unrelated if only content is taken into account.

Meta-content following protocols such as Friend of a Friend (FOAF) could, in principle, be also used as network arcs, but its use is not widespread, and it represents simply a binary relation (either you are a FOAF or you are not), while links have some quantitative quality (linking several times is different from linking only once).

In this work, links have been chosen over content because they are easily parseable from the document source; this choice allows for a low-dimensional representation of each blog which will be represented by a vector with as many components as blogs in the group under study. This obviously only holds if the number of relevant sites is smaller than the vocabulary needed to represent the same sites in a vector space model. It is also univocal: a link clearly identifies origin (the weblog it has been found in) and destination (from the URL). Links represent a real relationship among the blogs they join: they imply that, at least, one has read the other, which shows a kind of *community* relation. This is inferred because communities are created by reading, writing about other blogs or commenting on them. It is true that there might be other members of the community not uncovered by this method (for instance, loyal readers or people who use comments to participate); similarly, a member of the community could be linked to another via a blog not belonging to the set of blogs under study (Blogalia, in this case); however, we do not attempt to say the last word about

community structure in the blogosphere (as is usually called the set of all weblogs). Our aim is to portray a method to identify communities by considering hyperlinks a good enough indicator of community relationship.

Content (distance in vector space) or links (number of links, or just the existence or not of links) are used to create a complex network of the set of sites under study; consequently, a community must be defined by some measure that distinguishes, or makes apart, some sites from others. There are several possible network structures that could be considered communities: *cliques*, or sets of sites that link to each other, *bipartite cliques*, sets of sites which all link to another, different, set of sites, *k-cores or factions*, sets of sites connected to, at most, *k* other sites in the group, or *bipartite cores*, which includes both the connector and the connected sites. Most of these structures can be computed and displayed with programs such as Pajek (available at http://vlado.fmf.uni-lj.si/pub/networks/pajek/) or UCINET (available from http://www.analytictech.com/), but require some initial parameters such as the number of cliques or the number of cores we want to divide the original set into. All of these are valid definitions, and can be used in some cases. However, some of them are restrictive in the sense that they only take into account binary relations, and not the link weight (number of times it has been used) or direction. In the case at hand, direction is important: usually, some blog that has been ``pointed to'' might not even be aware of it. The majority of the concepts defined above do not create clear visual image of the community they are describing.

Sometimes, further steps must be taken to infer complex network communities. Some of them are geared toward specific communities, e.g. communities expressed via web pages or email messages, like the one we are dealing with in this paper. Gibson et al. proposed one of the first algorithms to infer web communities; it defined a community as a core of central, *authoritative* pages linked by *hub* pages. However, this definition is a bit fuzzy and does not provide clear-cut partitions of a set of websites, but it is interesting in the sense that it was one of the first to realize the importance of communities on the web, and to propose an algorithm to define them. Shortly afterwards, Flake et al. use a maximum flow/minimal cut algorithm to define the edges and nodes that act as boundary between communities.

There exist other algorithms that detect partitions of the original set according to properties of links, as opposed to properties of nodes. One of these is the Girvan-Newman algorithm (see Merelo et al. 2004 or Radicchi et al. 2003 for the state of the art and references), which detects links that, when removed, isolate some part of the original set. Clusters, or communities, are then computed according to where these removed links are. This algorithm discovers communities quite efficiently, but, once again, it does not discover the internal structure of each community, or the features that define them.

Recently, Radicchi et al. (Radicchi et al., 2003) review existing community definition and identification methods, claiming that most community definitions are algorithm-dependent, and propose a new definition for community discovery that is independent of the algorithm. Furthermore, they simplify Girvan-Newman algorithm by using purely local information to compute edge betweenness.

This paper, along with our previous work (Merelo et al., 2004), uses Kohonen's Self-Organizing Map (Kohonen, 1990), which is an unsupervised neural-network like algorithm that simultaneously performs clustering of input data, and maps it to a two-dimensional surface. Our objective is to demonstrate how the self-organizing map discovers underlying community structure efficiently, allows easy visualization of the complex network, highlights

the underlying topic that defines each community, and permits assigning new websites to a community by merely looking at its links.

The rest of the paper is organized as follows: first, we make a brief introduction to Kohonen's self-organizing map in the next section. The next section is devoted to present the results of applying Kohonen's self-organizing map to community discovery in Blogalia, and, finally, our conclusions are presented.

Kohonen's self-organizing map

Kohonen (Kohonen, 1990) originally proposed his self-organizing map inspired by previous work done by von der Malsburg as a model for self-organizing visual domains in the brain. Kohonen's SOM is composed of a set of *n*-dimensional vectors, arranged in a 2-dimensional array. Each vector is surrounded by other 6 (hexagonal) or 8 (rectangular arrangement) vectors. A size *n* neighbourhood of a vector is defined as the set of other SOM vectors whose index differs in less than a number *n*.

Kohonen's SOM, as many other heuristic methods, must be trained on the data it is going to model. Training proceeds by submitting vectors from the training set to the SOM, comparing distance from SOM vectors to it, and changing the closest SOM vector and the vectors in its neighbourhood to make them even closer to the input vector; self-organization emerges because different neighbourhoods, not the whole map, are updated every time a new vector is presented; and the learning proceeding in an unsupervised way. Other than that, SOM is similar to any other clustering algorithm, but, in this case, clusters are also arranged geographically. That is why it is said to perform a topographical mapping.

So far, the SOM has been used in many different applications, and, indeed, there there are several software packages, in different languages that implement it. The most popular is probably SOM_PAK (free and can be downloaded from http://www.cis.hut.fi/~hynde/lvq/.), which includes command-line programs for training and labelling SOMs, and several tools for visualizing it: `sammon`, for performing a Sammon projection of data, and `umat`, for applying the cluster-discovery Umatrix algorithm. We will use these programs in this paper to discover hidden communities in the Blogalia blog-hosting site.

Mapping weblog communities

The working set of websites that will be used in this work corresponds to weblogs hosted by Blogalia (http://www.blogalia.com/); it hosts around 200 weblogs, of which only162 actually link or are linked by other weblogs; these are the ones actually used in our study. All stories, and just the stories (excluding information in page templates, or dynamic newsfeeds, for instance) published in Blogalia up to September 2003 were used; there were around eleven thousand, which included around seventeen thousand links. Of those, roughly a quarter were links to other members of the community; this set of links will be used in this work to try to understand the Blogalia community structure. Each weblog is represented by the set of output links to other members of Blogalia. Of course, and due to this decision, other websites or weblogs are not considered, which means some sites closer to some blogs hosted in Blogalia than most of the inhabitants of that site might be ignored; however, in this paper, our intention was to discover communities within Blogalia, not all communities that included webs hosted by Blogalia.

In this work, each blog is represented by a vector whose components are the number of times it links to others in Blogalia; if a blog such as http://fernand0.blogalia.com/ links to http://atalaya.blogalia.com 7 times, the corresponding element will hold the value 7. Incoming and outgoing links are considered separately.

As a first approach to visualizing of the Blogalia social network, we used Pajek to plot input cores in this set. This graph revealed a densely connected cluster around the center of the figure, and a sparsely connected periphery composed mainly of abandoned or test sites. The densely connected core represents the most pointed-to blogs within Blogalia, corresponding, indeed, to *authorities*. Indeed, they have a high authority degree as computed by UCINET.

UCINET was also used to compute *factions*, that is, set of blogs which all point to the same blogs. The number of factions was preset to 3. In this case, the first faction corresponds roughly to the densely connected cluster revealed by the Pajek plot mentioned above; the third, to the sparsely connected group of blogs, and the second, to all the blogs in between.

The same data was analysed using Kohonen's self-organizing map. The software used was SOM_PAK version 3.2. From the links array, two different analysis were performed: by rows and columns. Rows represent the set of blogs every blog links to, and columns represent the set of blogs that links to a particular one. That means that SOM was applied to blogs represented by *incoming* and *outgoing* links. On each map, Umatrix analysis was applied: this analysis shows how the set is clustered, so that clusters tend to stand out.

Different results have been obtained by training representing blogs by incoming or outgoing links. In the first case, a single block, containing the most usually linked-to blogs, stands out. This block roughly corresponds to the core mentioned above, and the first faction obtained by faction analysis. The scenario that uses outgoing links is shown in figure 4 is a bit more discriminating, but, once again, distinguishes factions and cores as computed by other methods.

But it would also be interesting to look at what makes blogs cluster together in a single node, or what they have in common; it results that, usually, they all point to one or two nodes, which become local *authorities;* that means that the blogs mapped to a single node roughly correspond to bipartite cliques, that is, set of nodes whose link pattern is similar.

To infer communities from this map, a first approach would then be to assign a community to each node, which would yield several dozens of communities out of the original hundreds of websites. This is not satisfactory, however, for two reasons: first, nodes which are closer in the Kohonen map might also belong to the same community, and second, some of the blogs that are mapped to a single node do not actually belong to any community. Consequently, we will have to take, a second approach, based on the usual clustering techniques applied to Kohonen maps postprocessed with the UMatrix algorithm: clusters are ``white'' zones surrounded by ``black'' boundaries; white zones represent nodes that are close to each other, while black nodes are far apart from those around it; in this case, a single community can be appreciated.

Since this can be only identified by visual inspection, a new definition of community cannot be deduced, specially in this case when there is not a clear-cut division in two or more clusters. So we will introduce a new definition of community as *the set of network nodes that fall on the same node of a self-organized map*. This definition is functional, and, besides, allows assignment of new nodes just by taking into account its links to the members of the

set under study. An additional advantage is that navigation from a community to another is possible, just by moving from a node to its neighbours on the Kohonen map. Besides, a single representative for each community can be extracted from each node on the network.

There is indeed some congruence with communities defined this way and other concepts. If we map the factions computed by UCINET to the map, it will result that most nodes are occupied by blogs that belong to a single faction, while some nodes, acting usually as a boundary, are shared by several factions.

The same tool can be used to map the evolution of a set of (generally linked) weblogs. Instead of using the state of the blogs at a particular point in time, several snapshots are taken at different intervals, taking into account the accumulative links created during that period. The first *snapshot* would include the first 3 months, the second would include those links along with the links created during the subsequent 3 months, and so on. The Kohonen map, in this case, is trained with the set of all vectors for all periods, and, when training is finished, the vectors representing each blog for each time period are mapped to it.

The picture that arises shows that time evolution is reflected on the Kohonen map in the following way: blogs are mapped during the first period to a very concentrated set of neurons; but, with time, they start to spread away from it, moving towards the borders and corners of the map. Finally, when the last time period is mapped, “mature” blogs (those that were created first, and those that have the highest standing within the community) are placed along corners; while newly created blogs are still in their primeval quarters in the starting cell.

Besides, this map can be used to find out what is the “development stage” of a new blog, once it has been trained. Trajectories of particular blogs can also be plotted on the map: for most of them, they start to evolve for some time periods, until finally they settle. This might coincide with exploration/exploitation periods of the blog authors; but, of course, this would have to be studied further.

Conclusion

Web content creation has undergone lately, under the influx of easy content-management programs such as weblogs, an extraordinary expansion, which, so far, shows no sign of abating. Interest groups are created spontaneously among web users, and it is enlightening to study and identify these groups from the sociological, economical and technological point of view. Since web-community formation is generally spontaneous, without an explicit register or inscription by those that integrate them, and, besides, a particular website might belong to several communities, one of the first problems posed by its study is its identification and representation.

In this paper, we give some details on using a technique well known in the pattern recognition and data mining fields: Kohonen's self-organizing maps; our approach was originally presented in (Merelo et al., 2004). As has been shown in this paper, communities identified by analyzing self-organizing maps using UMatrix are on a par with those identified using other techniques, such as faction analysis or core extraction, with the additional advantage that *community navigation* can be achieved by using the map: blogs on the same node, or adjacent nodes, belong (in a fuzzy sense) to the same community. The self-organizing map, besides highlighting the different communities and groups present on the sample, make an useful visual representation.

Acknowledgements

This paper has been funded in part by project TIC2003-09481-C04-04, of the Spanish ministry of science and technology, and a project awarded the Quality and Innovation department of the University of Granada. Fernando Tricas is with the Group of Discrete Event Systems Engineering (GISED), and his work is partially supported by TIC2001-1819, financed by the Spanish Ministerio de Ciencia y Tecnologia. We are also grateful to Víctor Ruiz for his creation and continuing support of Blogalia, and its members, for their encouragement during the realization of this work.

References

(Kohonen, 1990) Kohonen, T., *The Self-Organizing Map*, Procseedings IEEE (1990), 78,1464 ff..

(Merelo et al., 2004) Juan J. Merelo-Guervós, Beatriz Prieto, Alberto Prieto, Gustavo Romero, Pedro Castillo-Valdivieso, and Fernando Tricas. *Clustering web-based communities using self-organizing maps*. In Kommers et al. [129], pages 158-165. Available from http://geneura.ugr.es/jmerelo/papers/72.pdf.

(Radicchi et. al., 2003) F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, *Defining and identifying communities in networks*, Arxiv CondensedMatter, abstract cond-mat/0309488, September 2003, available from http://es.arxiv.org/abs/cond-mat/0309488.

(Tricas et al., 2003) Fernando Tricas, Juan J. Merelo, and Víctor R. Ruíz. *Do we live in a small world? Measuring the spanish-speaking blogosphere*. In Thomas N. Burg, editor, Blogtalks, Proceedings of BlogTalk A European Conference on Weblogs, Viena, Austria, May 23-24, 2003, pages 158-173, May 2003. Available from http://www.blogalia.com/pdf/20030506blogtalk.pdf.