Visualizing the evolution of a web-based social network

Beatriz Prieto\textsuperscript{a}, Fernando Tricas\textsuperscript{b}, Juan J. Merelo\textsuperscript{a,\textdagger}, Antonio Mora\textsuperscript{a}, Alberto Prieto\textsuperscript{a}

\textsuperscript{a}Departamento de Arquitectura y Tecnología de Computadores, Universidad de Granada, Granada, Spain
\textsuperscript{b}Departamento de Informática e Ingeniería de Sistemas, Universidad de Zaragoza, Zaragoza, Spain

Received 16 March 2007; received in revised form 25 October 2007; accepted 1 November 2007

Abstract

Weblogs are dynamic websites updated via easy-to-use content management systems and organized as a set of chronologically ordered stories, frequently built around a link or including links to other weblogs. Since they are managed by individuals, their links tend to mirror or, in some cases, establish new types of social relations, thereby creating a social network. Studying the evolution of this network allows the discovery of emerging social structures and their growth trends. In this paper, we demonstrate the advantages of using the self-organizing maps (SOM) to visualize the evolution of a social network formed by a set of blogs, from their beginning to their current state. By observing the position a weblog is mapped to, it is easy to see what communities it belongs to nowadays, and how and when it became a part of those communities. The proposed procedure gives some insight on how communities are formed and have evolved. In this study, we apply this method to Blogalia, a blog-hosting site from which we have obtained a complete set of data and, by using SOM projections, we have drawn some conclusions on what drives the evolution of its implicit social network.

Keywords: Weblogs; Blogs; Neural networks; Self-organizing maps; Application of Kohonen maps; Web-based communities; Social networks; Social network evolution; Social network visualization

\textdagger\textsuperscript{Corresponding author.}

\textit{E-mail addresses:} bprieto@atc.ugr.es (B. Prieto), ftricas@unizar.es (F. Tricas), jmerelo@geneura.ugr.es (J.J. Merelo), amorag@geneura.ugr.es (A. Mora), aprieto@ugr.es (A. Prieto).

© 2007 Elsevier Ltd. All rights reserved.

doi:10.1016/j.jnca.2007.11.003
1. Introduction

According to the Tech Encyclopaedia (TechEnciclopedia), a weblog is a web page that contains links to other web sites that cover a particular subject or meet some other type of criterion. However, this definition does not cover the current situation of these kinds of websites. According to Wikipedia, the free encyclopaedia (Wikipedia), a weblog—also known as blog—is a website where stories by one or more authors are chronologically compiled, and which have some common elements, such as a list of links to other weblogs, an archive of past stories (also called posts), and permanent links (permalink links) so that anyone can mention (by linking) an annotation, or a function that allows the addition of comments. The difference between weblogs and mainstream news sites (which is becoming increasingly fuzzy nowadays) is that blog posts usually include links to other web pages (not necessarily a weblog) as a reference or in order to extend the information; they are also, in most cases, personal as opposed to company-owned. Therefore, links reflect, in some way, personal relationships constituting thus the edges of a social network whose nodes are blogs, and whose actors are blog writers.

Weblogs are gradually acquiring greater importance by becoming an alternative source of information and opinions. Several million have been created worldwide: according to Perseus (The Blogging Iceberg, 2004), 4.12 million blogs had been created just through BlogCity, BlogSpot, Diaryland, LiveJournal, Pitas, TypePad, Weblogger and Xanga, the most popular blog-hosting sites back then. Closer to home, the growth in the Spanish blogosphere in the last few years has been spectacular. Since the creation of the “Blogometro” (Merelo et al., 2005), that visited thousands of Spanish blogs daily and generated a list with the most linked URLs, the number of Spanish blogs has grown to several millions.

A weblog by itself is not necessarily important, but as part of a community (Newman, 2003), its importance cannot be disregarded. Social networks are a type of complex networks (Albert and Barabasi, 2002) wherein nodes are usually social entities (individuals or groups), and links represent relationships among them. The social network created around a weblog is formed by the authors or editors of the weblogs, the people who send comments to their posts, and quiet but frequent readers who may also have their own weblog. However, only a part of this community can be measured: the subset formed by weblog authors. All weblogs in the world can be seen as components of a set of communities, each one with its own axioms, hierarchies, idols and enemies. 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 accounted-for weblogs in the Spanish-speaking community (Merelo et al., 2003) at the time of that study) are connected to each other by a finite set of links, that is why it is not so interesting to segment the social networks in groups as to visualize the structure of the graph and how it became to be the way it is. A visualization tool, such as the one we propose in this paper, will help to place blogs within the communities they belong to, and since it has been created around a map, it also creates a tool for mapping new, so-far unknown blogs, to the place they occupy within those, already known, communities.

In this paper, we present a procedure that makes it possible to highlight the collective behaviour of a set of blogs, using the facilities provided by the self-organizing maps and U-matrices.

The rest of this paper is organized as follows: firstly, we sketch the state-of-the-art (Section 2) in visualization of the evolution of social networks and communities; followed
by a brief introduction to self-organizing maps (Section 3). Section 4 is devoted to the implementation of the application using the SOM Toolbox. Section 5, experimental procedure, presents the evolution of weblog communities by applying the self-organizing maps (SOM) to the different stages of Blogalia; whose interpretation is made in Section 6, visualizing the evolution of weblog communities. Finally, the conclusions can be found in Section 7.

2. State-of-the-art

The last few years have seen an explosion in the study of social networks and their evolution, and, more specifically, of blogs as complex social networks. Let us look at these two issues separately.

The evolution of complex networks has been studied mainly from a theoretical point of view by formulating hypotheses, such as the preferential attachment hypothesis (Barabasi and Albert, 1999), which states that social networks grow mainly by following a rich-get-richer scheme (or linked-gets-more-links, in this case). This hypothesis was later corrected and improved by Geng and Li (2005) by taking into account links among already-existing nodes and by rewiring and deleting links (the good-get-richer hypothesis) (Caldarelli et al., 2002). However, in the real world, it is sometimes not as important to know that the good get richer (in this case, get more inbound links), but to see why they are considered better than other nodes, which may seem quantitatively very similar (for instance, in the case of blogs, similar age, or similar posting frequency). And, in many cases, the issue is to know where a social network comes from, in order to find out where it might be going in the future. It could be said that, since every link is the result of a rational decision, it is impossible to make any prediction on the future of a social network, such as the one we are dealing with. However, that decision has some constraints, which make the behaviour of a set of blogs collectively predictable and, in this paper, we attempt to initially highlight their behaviour through the use of a visualization tool.

The visualization of complex networks is mainly carried out by using programs such as Pajek (Batagelj and Mrvar, 2003) or NetDraw. The former is able to represent time-stamped information, mainly by keeping old nodes in place, and by trying to fit in new nodes and their connections. However, it is not specifically suited for this purpose. This is why new tools have been proposed for the explicit representation of time-evolving complex networks. For instance, Gloor et al. (2004) propose the WEBINAR system, which simultaneously illustrates the evolution of several macroscopic values along a time window. Some other authors (Holme et al., 2004) have also studied the evolution of a real-world social network, namely, an online dating network, focusing mainly on the correlation of different kinds of links (guestbook entries and IM messages, for instance), and the overall size of the network and the number of links, without paying attention to its visualization. Their main result is the assortative/dissassortative nature of the network. Their work can be used to calibrate theoretical models such as the ones mentioned above or those exposed in Dorogovtsev and Mendes (2003).

This paper does not tackle directly the problem of community detection, rather focusing on visualization of relative community membership. However, there is an extensive body of work carried out on the detection of communities without studying the dynamics.

---

Girvan and Newman 2002; Newman and Girvan, 2004), for instance, propose algorithms mainly oriented to produce fast solutions. In the work performed by Guimera et al. (2003) the authors explore the use of these detection algorithms to improve the knowledge on organizations. Kumar et al. (2005) and Broder et al. (2000) reported the detection of communities via co-citation; that is, sites that appear on pages together can be related, even if they do not give a reference to each other.

The work presented in this paper has the same orientation: it intends to offer a procedure for visualizing the evolution of a complex network along time, that gives some insight on which mechanisms are functioning during that evolution. That is, the method presented in this paper advances the state-of-the-art by presenting a procedure that shows the evolution and current state of a social network using a computationally light method such as Kohonen’s self-organizing map, which allows to extract meaningful information about community relationships and how they came to be that way. Kohonen’s SOM, by using just information on linking patterns, and is able to map thematically and chronologically related blogs to nodes that are close to each other. Moreover, by using the linking patterns along time, the evolution of a blog from a part of the map to another can also be easily visualized.

3. Kohonen’s self-organizing feature map

In this section, we make a brief introduction to the self-organizing feature map (SOFM) and a general description of the procedure to generate and implement it.

It is usually admitted that in humans the different sensorial inputs (motor, somatosensorial, tactile, visual, auditory, etc.) are mapped, topologically ordered, in specific areas of the brain cortex, forming authentic cerebral computational maps. According to this simple model, neurons transform input signals into a distribution of probability coded by its position. This distribution represents the computed values in positions within a map where relative maxima of the activity are obtained (von der Malsburg, 1973; Knudsen et al., 1987).

Kohonen (1982, 2001) asserted that the origin of the cerebral maps could be found in the search for an economic representation of the data and their interrelations. This fact constitutes one of the central objectives of information sciences, and this author proposes a model of artificial neural networks called SOM (SOM, also known as SOFM), also referred to as Kohonen maps.

Although the main structures of the cerebral networks are genetically fixed, the dependency of the sensorial projections with this evolution has been experimentally proven in the recovering evolution of sensorial handicaps in children or people with mutilations in the nervous tissue. In fact, a self-organization caused by the sensorial information takes place in such a way that, subsequently, the location of the cell in the network where an answer is obtained gets to be specific to a certain characteristic in the input data set. Since this type of network is usually distributed in a plane (two-dimensional structure) it can be concluded that the projections preserve the topologic relations while simultaneously creating a dimensional reduction of the representation space. In particular, the main objective of the SOM (Kohonen, 1990) is to transform a set of input patterns of arbitrary dimension into a discrete map, usually of two dimensions, and to adaptively make this transformation in a topologically ordered way.
Fig. 1 illustrates a scheme of a two-dimensional neuronal network, usually used to generate a discrete map. Note that each neuron of the network is completely connected to all the nodes of the input layer. The network represents a feed-forward structure with only one computational layer formed by neurons or model vectors (also denominated prototype vectors, or weight vectors), \( \mathbf{m}_i \in \mathbb{R}^n \), arranged in rows and columns. Sometimes a one-dimensional model is used and can be considered as a particular case of the one in Fig. 1, considering only a row or a column.

Each input pattern presented to the network is typically represented by a located region or “point” of activity produced on a stable background. The position and nature of this point usually changes with the input pattern. All the neurons of the network would have to be exposed to a sufficient number of input patterns to ensure that the self-organization process is suitable.

There are four essential processes involved in the formation of self-organizing maps:

1. Initialization of model vectors. It is common to use one of the following alternatives: (1) assigning small random values (random initialization), (2) initializing with random samples taken from the input data set (sample initialization), (3) regularly distributing the model vectors in the plane (regular initialization), and (4) selecting the model vectors in an orderly form along the linear subspace spanned by the eigenvectors corresponding to the two largest principal components of input data (linear initialization).

2. Competitive process. For each input pattern, \( \mathbf{x} \in \mathbb{R}^n \), a discriminant function is evaluated; its value provides the basis of a process of competition between the different neurons. The winner of the competition is the neuron that produces the greater value of the discriminant function. In order to find the best similarity of the input vector \( \mathbf{x} \) with the \( l \) model vectors \( \mathbf{m}_i \) that conform the network, the scalar products, \( \mathbf{m}_i \cdot \mathbf{x} \) for \( i = 1, 2, \ldots, l \), will be compared, and the greatest one will be selected. The neuron, \( \mathbf{m}_c \), with the greatest product (which mathematically means that the Euclidean distance between the vectors \( \mathbf{x} \) and \( \mathbf{m}_i \) is minimum) determines the centre of the topological neighbourhood, \( h_{cd} \). It will be verified that:

\[
||\mathbf{x} - \mathbf{m}_c|| = \min_i \{||\mathbf{x} - \mathbf{m}_i||\}, \quad i = 1, 2, \ldots, l.
\]
This expression symbolizes the essence of the process of competition between the neurons of the network. The neuron that satisfies (1) is said to be the one of maximum similarity or the winner neuron or the best-matching unit (BMU), and the effect of the process is that the space, $X$, continuous of activation patterns is projected in a discretized output space, $M$.

3. **Cooperative process.** The winning neuron determines the spatial position of the centre, $c$, of a topological neighbourhood, $h_{ci}$, of the excited neurons. In this way, the basis for the cooperation between the neurons of that vicinity is obtained.

The neurons of the map can be arranged on a rectangular or a hexagonal lattice (see Fig. 2) in the two-dimensional case. The innermost polygon corresponds to the 1-neighbourhood, the second to the 2-neighbourhood and the biggest to the 3-neighbourhood. The *neighbourhood function* determines how strongly the neurons are connected to each other.

Let us define the lateral distance, $d_{ci} = ||r_c - r_i||^2$, between the winning neuron, $c$, and a generic neuron of the network, $i$; and $h_{ci}$ a neighbourhood function centred in the winning neuron, which determines the degree of influence of the winning neuron on neuron $i$. The neighbourhood function must be unimodal with the lateral distance, $d_{ci}$, if a similar behaviour to the neuro-biological one (where cooperation through lateral interactions takes place) is sought. In this way, the following requirements must be satisfied:

- $h_{ci}$ has to be symmetrical around its maximum in $d_{ci} = 0$; that is, the maximum is obtained in the winning neuron, since the distance is zero.
- The amplitude of $h_{ci}$ must decrease monotonically when the lateral distance, $d_{ci}$, increases; tending to zero when $d_{ij} \to \infty$. This is a necessary condition in order to obtain convergence.

A neighbourhood function which satisfies the previous requirements and which is frequently used, is the *Gaussian function*:

$$h_{ci} = e^{-d_{ci}^2/2\sigma^2}. \quad (2)$$

The parameter $\sigma$ defines the effective width of the neighbourhood function, or the degree in which the winning neuron affects its neighbourhoods.

While the inputs arrive at the network in successive moments of time, $t = 0, 1, 2, \ldots$, a regression process of the model vectors, $m_i$, into the space of observation vectors, $x$, occurs. During this process, the effective width, $\sigma$, must be reduced monotonically with the regression step, $t$, so that the neighbourhood function (2) becomes:

$$h_{c(x),i}(t) = e^{-d_{ci}^2/2\sigma(t)^2}. \quad (3)$$

4. **Adaptation process.** The adaptation tries to increase the individual parameters of the winning neuron in relation to the input pattern. The adjustment takes place in such a way that reinforces the answer of the winning neuron with the application of similar input patterns. The adaptation specifies the way the regression of the model vectors into the space of observation vectors must be carried out, and it is usually done according to the following model (Kohonen, 1982, 2001; Haykin, 1999):
where \( \alpha(t) \) is a function which decreases monotonically with the regression step, \( t \), known as the learning rate.

The previous expression (4) is applied to all the neurons of the network. However, the neighbourhood function, \( h_c(x,i) \), causes a change of weights only in the neighbourhood (the value of \( h_c(x,i) \) is null and, therefore, the weights are not modified in the neurons, \( i \), far from the winning neuron, \( c \)). The application of expression (4) has the effect of moving the vector of weights, \( m_c \), spatially from the winning neuron \( m_c \) towards the input vector \( x \). Thus, after repetitively applying the learning sequences, and due to the update of the neighbourhood function, the model vectors tend to follow the distribution of the input vectors. Consequently, the algorithm leads to a topologic arrangement of the characteristic map of the input space, in the sense that adjacent neurons in the network tend to have similar weights vectors.

Once the SOM algorithm has converged, the map of characteristics, \( \Phi \), obtained in the output space, \( M \), displays relevant statistical properties of the input space, \( X \). The following ones could be emphasized (Haykin, 1999):

- **Approximation of the input space.** The feature map, \( \Phi \), represented by the set of model vectors \{\( m_i \}\} in the output space, \( M \), gives a good approximation of the input space \( X \).
- **Topological order.** The feature maps, \( \Phi \), obtained through the SOM algorithm are topologically ordered in the sense that the spatial localization of a model vector in the network corresponds to a particular domain or characteristic of the input patterns.
- **Density matching.** The feature maps, \( \Phi \), reflect the variations in the statistical distributions of the inputs: the regions of the input space, \( X \), such that the sample vectors \( x \) are generated with high probability, are projected in large domains of the output space, \( M \), and, therefore, have a better resolution than the regions corresponding to the zones of the input space, \( X \), in which the input patterns \( x \) are generated with a smaller probability.
- **Feature selection.** When given a set of data of the input space with a non-linear distribution, the self-organization map is able to select a set with the best features in order to approximate the underlying distribution.

The main applications of the self-organizing map are: visualization, clustering, interpolation or function modelling, classification (in some cases) and vector quantification. The SOM effectively creates a map of the input space that can be used, once trained, to associate new inputs to the features of vectors used for training. In this case, the feature we are interested in is the relationship among blogs, the natural groupings created among them, and what lies behind that relationship (be it thematic, political or personal proximity).

4. Implementation using the SOM Toolbox

There are many software packages that implement SOM, such as the SOM Toolbox for Matlab, or the SOM_PAK created by Kohonen’s team (Kohonen et al., 1996; Alhoniemi et al., 2000). This package contains a series of functions which make it possible to train SOM with different network topologies and learning parameters, calculate different errors, quality and measures for the SOM, visualize SOM using \( U \)-matrices, component planes,
cluster colour coding and colour linking between SOM and other visualization methods, and to carry out a correlation and cluster analysis using SOM. The U-matrix is generated by using a post-processing method (Ultsch, 1993); Fig. 3 highlights the different clusters as well as the metric-topological relations between input vectors and outstanding variables.

The algorithm and parameters used are described below.

**Step 1. Initialization of model vectors**

In the application described in this paper, a linear initialization has been used, so the SOM Toolbox selects a regular array of vector values positioned on the subspace spanned by the eigenvectors corresponding to the two greatest principal components (eigenvalues) of the input data. Usually, the number of neurons chosen is as big as possible, but the computational load increases quadratically with the number of neurons. In this study, the default number, \( l \), of neurons considered by the toolbox is 147. The package obtains an approximate default number of neurons by applying the following expression:

\[
 l = 5 \cdot \sqrt{n},
\]

where \( n \) is the number of training vectors (the total number of training vectors used is \( n = 810 \)).

A two-dimensional shape is used; according to the recommendations of the toolbox authors, length along one dimension is longer than the other, which helps the map to orientate itself properly. Particularly, the map created by the toolbox has \( 21 \times 7 \) neurons.

A hexagonal lattice (Fig. 2b) has been selected so that all six neighbours of a neuron are at the same distance (as opposed to the eight neighbours in a rectangular lattice, Fig. 2a), avoiding topological distortions. In this way, the maps become smoother and more pleasing to the eye.

**Step 2. Sampling**

In each training step, \( t \), one sample vector \( x(t) \) from the input space, \( X \), is chosen at random.

**Step 3. Competition**

The objective is to find the (winning) neuron, \( m_c \), which is most similar to the sample \( x(t) \), in step \( t \), using the criteria of minimum Euclidean distance, by applying expression (1).

![Fig. 2. Bi-dimensional network that generates a SOM and post-processing to obtain the U-matrix.](image-url)
Step 4. Cooperative and adaptation processes

The synaptic weight of every neuron is updated by using expression (4), where the neighbourhood function, $h_{c(x),i}$, centred in the winning neuron $c(X)$, is given by expression (3). The learning rate function used, $\alpha(t)$, is the inverse function with an initial value of 0.5. The toolbox uses a large neighbourhood, a big initial value for learning coefficient and a short training.

Step 5. Continue

Go to step 2 (sampling) until a non-significant variation of the model vector occurs. Generally, the number of training steps should be at least 10 times the number of the map units. Since all the samples $x(t)$, $t = 1, \ldots, 810$ (162 times five periods) are available, a batch version of the algorithm is used. A significantly greater speed is obtained by using this algorithm, and it sometimes obtains even better results than the incremental approach. Particularly, the following computational scheme has been used (Kohonen, 2001):

Step 1: Initialize the model vectors $m_i$. Do $i = 1$ and $t = 1$. 
Step 2: Select the map unit $i$.
Step 3: Collect the most similar observation samples, $x(t)$, to the model vector that belongs to the neighbourhood set of unit $i$.
Step 4: Replace the model $m_i$ with a weighted average over the samples collected on Step 3:

$$m_i(t+1) = \frac{\sum_{j=1}^{162} h_{c(j)}(t) \cdot x_j}{\sum_{j=1}^{162} h_{c(j)}(t)}.$$  (6)

Step 5: Do $i = i + 1$ and go to Step 2.
Step 6: Repeat 10 times from Step 2.
Step 7: Visualization of the results. End

5. Experimental procedure

The working set of websites used in this study corresponds to weblogs hosted by Blogalia (http://www.blogalia.com/), which hosts around 200 weblogs. Only 162 of these
weblogs link, or are linked by others; these are the ones used in our work. Only the stories or posts (excluding information in page templates, blogrolls, or dynamic news-feeds, for instance) published in Blogalia were used for this study; there were around 11,000, which included roughly 17,000 links. Of those, approximately a quarter were links to other members of the site. This set of links is what will be used in this study in an attempt to understand the structure of the Blogalia community. Each weblog is represented by a set of outgoing links to other members of Blogalia, which logically do not include non-Blogalia websites or weblogs. This means that a number of sites, which are closer to some Blogalia weblogs than current Blogalia members, may be ignored. However, in this paper our intention is to map the evolution within Blogalia, and not in all communities that included webs hosted by Blogalia. There are several reasons for this: firstly, it is easier for us to work with a manageable subset of the blogosphere; secondly, two of this paper’s researchers keep our blogs at Blogalia, making it easier to understand and validate them.

Data has been obtained directly from the Blogalia database, by extracting just the text corresponding to the stories (excluding templates and other stuff, such as comments), and scraping hyperlinks from them by using a Perl program, and is formed by 162 components containing 162 2-vectors. Each component represents the number of times a blog links to another. For instance, in Fig. 4 the component “100” means that the blog \( i \) has linked the blog \( j \) 100 times, that is, the component \((i,j)\) of the vector is 100. Otherwise, it can be interpreted that blog \( j \) is linked 100 times by blog \( i \). Therefore, the columns represent incoming links and rows represent outgoing links.

In this work, incoming links have been considered as well as outgoing links. There are five groups of data (five groups of 162 × 162 vectors) each one corresponding to the links of each blog after every 4-month period (except for the first period, which is 8 months long). That is, the first group contains the links to the blogs for the first 8 months, the second one, links to the same blogs accumulated by 4 more months (12 months), and so on. In order to visually distinguish the data corresponding to each period, every vector has been labelled with the blog name (the first part of the URL, for the sake of simplicity) preceded by a number 1, 2, 3, 4 or 5, according to the corresponding period. For example, \( 1atalaya \) will be the \( atalaya \) vector for the first period; \( 2atalaya \) will designate the \( atalaya \) vector for the second period, and so on, all of them corresponding to \http://atalaya.blogalia.com/\ (this is J.J. Merelo’s blog, one of the authors of the paper). Since we have a static view of the database, which is not changed during data extraction, vector values can only become bigger in each period; links do not disappear.

![Fig. 4. Organization of prototype vectors.](image-url)
In fact, this representation of linking patterns puts a soft limit on the size of the networks that can be approached using this technique. Even though there are no theoretical limits, SOM loses efficiency when dimensionality increases, due to the topology of high-dimensional spaces. Logically, this also results in an increase in the computation effort. In this case, an analysis to extract the main nodes should be performed, which could be something as simple as choosing the top $n$ nodes ordered by inbound links. The experimental procedure is established through the following steps:

1. Reading the normalized data set. All the data (the five groups, previously normalized) has been used to train the SOM.
2. Creating, initializing, and training the SOM, using the parameters exposed in Section 4.
3. Once the output map was obtained, each group of data corresponding to each period was projected independently into the SOM in order to visualize the evolution of the network.
4. In order to better understand the map, the vector labels (that correspond to the blog names) of each group under the SOM map have been shown.

Thus, the methodology we propose to visualize weblog communities and its evolution can be expressed this way:

1. Select a group of weblogs linked by some characteristic (same physical hosting, for instance), with a soft limit of around 500 blogs. If there are more than that amount of blogs, choose only the top 250–500 blogs with most incoming links.
2. Extract links contained in stories, and take only those links to the members of the community (or the selected group within that community).
3. Slice periodically the links, and choose the period so that the change in the accumulated number of links is significative. Four months is probably right for many cases, but in some type of communities, a bigger value could be needed.
4. Represent each blog by the vector of links to the rest of the blogs, considering a vector for each time period, and train a Kohonen map on that set of vectors following the procedure described in Section 4.
5. Extract some characteristics of some of the blogs (such as age, political tendencies, topic), ideally the most linked-to, and label zones of the map according to age and these characteristics, as shown in Section 6.
6. The resulting map permits to visualize those features of blog which were unknown in advance, how close they are to the rest of the blogs, and will also be used as a diagnostic tool for new blogs just by representing them with the set of links to the blogs used to train the map. This is also shown in Section 6.

6. Visualizing the evolution of weblog communities

As has been previously indicated, a self-organizing feature map is characterized by the formation of a topographic map of input patterns (weblog vectors) in which the special locations (i.e. coordinates) of the neurons in that lattice correspond to intrinsic features of the input patterns. In the context of this paper, in a self-organizing way, blogs with similar link characteristics will be next to each other in the output representation layer. That is
obtained through the minimization of distortion (quantization map), but also of
topographic error, which guarantees that the topographic organization is the input space
is kept in the output space. For the maps shown in the paper, quantization and
topographic errors hover around 0.01.

The projection of each group of incoming links into the SOM is illustrated in Fig. 5 (the
first period in Fig. 5a, the second one in Fig. 5b, etc.). The size of the black hexagons is
proportional to the number of times each unit of the map has been the BMU (best-
matching unit) of the data set.

Fig. 5f illustrates the clustering structure (U-matrix) of the SOM, obtained by
minimization of the quantization of topographic error. In this map, topographic error is
0.007 and quantization error 0.013. Indexed colour representation is used to show the
colour coding of the clustering. The main feature in clusters seem to be age or number of
links; but not all blogs with a certain age have a high number of links; only those with a
certain position within the community. Besides, high-authority (high number of links)
blogs are divided in two clusters (left-hand side and right-hand side of the map), with the
left-hand side occupied by blogs that were among the first in Blogalia, and with a certain
political trend, and the right-hand side by more technologically oriented blogs.

It is possible to observe that initially (Fig. 5a) most blogs are in the upper right corner
(the biggest black hexagon). Each blog in this hexagon has null values in its components,
that is, these blogs have no incoming links. As time passes the number of blogs in this
hexagon decreases as each blog begins to receive more links. Blogs start to move
throughout the map in each generation as they start to receive links. In Fig. 5b–e we can
observe how the space represented by the map is populated by different weblogs, while the
newest ones remain in the upper zone in addition to the blogs that are not very active (even
some that have been abandoned). On the other side, the most active weblogs descend and
appear at the bottom of the map. In Fig. 5a only three weblogs appear on the bottom row
(corresponding to the blogs that were first created). In Fig. 5b there are two more blogs in
this zone. In Fig. 5c the same blogs are at the bottom but in rows near them new weblogs
appear. In Fig. 5d there are three new weblogs on the bottom row and, finally, in Fig. 5e
we can observe that this situation has become stable. It might be more interesting to see
how the whole map becomes full of sites, having started off as almost completely blank.

In order to see this effect in detail, the blogs with more than 100 incoming links in the
fifth period have been selected, so its evolution along different periods of time is revealed.
Table 1 illustrates the number of links of the most linked-to blogs along the five periods
studied. Some blogs were created later than others. “Atalaya” was set up in May 2002, but
was only heavily used from period 2 onwards; “pawley” and “fbenedetti” were set up
during the third period, and so on. In general, 0 incoming links indicate that a blog was
still inactive.

The trajectory followed by each vector along the timeline is also marked on the map
(Fig. 6). The labels indicate the vector name and the corresponding period (1 represents the
first period, 2 represents the second period and so on).

The projections of vectors into the SOM map of the most linked blogs are moving
downwards and towards the lower edge. It is possible to notice the formation of three
groups based on the position that reaches each vector. Vectors representing the blogs
fernand0 (http://fernand0.blogalia.com/ is F. Tricas’ blog, the blog of another of the
authors) and atalaya, go to the bottom right-hand corner (Fig. 6). Vectors rvr and javarm
arrive at the bottom middle (Fig. 6). This cluster would be called “founding fathers” since
they are a few of the original blogs and the most heavily linked sites. Vectors fbenedetti, jkaranka, and pawley, new, but quite active, group, move to the bottom left corner.

Blogs that heavily link to each other also appear closer in the map, which indicates that they follow similar linking patterns, that is, they are linked by the same set of blogs. For example, atalaya (76 links to fernand0), www (14 links to fernand0), and jkaranka (12 links to fernand0), are all quite close to fernand0 as is shown in Fig. 7. Again, the black hexagons are proportional to the number of hits in each map unit.

Since we have separately considered incoming and outgoing links, the projection of each group of outgoing links into the SOM is shown in Fig. 8. Indexed colour representation is used to show colour coding of the clustering. The meaning of black hexagons is the same as

Fig. 5. The clustering structure: (a) first, (b) second, (c) third, (d) fourth and (e) fifth group of data (incoming links) projected into the SOM; the majority vote for each neuron is synthesized in the last graph (f). Final quantization error: 0.013. Final topographic error: 0.007. Although there is no clear trend (since the determining quantity is age, not period), the purple cluster has a majority of old-timers (periods 3–5); period-1 blogs show up mainly in the upper part of the map, and in borders.
previously. In Fig. 8a–e the first, second, third, fourth and fifth group of data respectively (outgoing links) are projected into the SOM.

Table 2 contains the blogs that have the most outgoing links. We have also chosen those with more than 100 outgoing links in the fifth period.

<table>
<thead>
<tr>
<th>Blog</th>
<th>Number of incoming links-to blog (&gt;100 links)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Period 1</td>
<td>Period 2</td>
</tr>
<tr>
<td>fernand0</td>
<td>29</td>
</tr>
<tr>
<td>atalaya</td>
<td>3</td>
</tr>
<tr>
<td>rrr</td>
<td>39</td>
</tr>
<tr>
<td>www</td>
<td>68</td>
</tr>
<tr>
<td>javarm</td>
<td>50</td>
</tr>
<tr>
<td>pawley</td>
<td>0</td>
</tr>
<tr>
<td>verbascum</td>
<td>30</td>
</tr>
<tr>
<td>fbenedetti</td>
<td>0</td>
</tr>
<tr>
<td>fkaranka</td>
<td>0</td>
</tr>
<tr>
<td>durnmith</td>
<td>0</td>
</tr>
<tr>
<td>paleofreak</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 2 illustrates some differences from Table 1. Those that appear in this second table are *jaio-la-espia* and *eledhwen*, which are blogs with many outgoing links. This indicates that they act as hubs, sending links to many other community members, but not getting them back in return. On the other side (blogs not included in Table 2) we find *www*, *daurmith* and *paleofreak*. The first one corresponds to the front page of the site, which was written by Victor Ruiz (the owner of Blogalia) to do some evangelism about blogging, and to advertise important news for the site (such as anniversaries). In this way, it is a blog that every member of the community is supposed to see when he or she visits the site. Everyone also expects that the others have read this site and may not feel the need to link to it. The second one, *daurmith*, is one of the oldest members of the site. We believe that the feeling about this blog is similar to the previous one: everybody is expected to read one of the oldest and most respected bloggers and, for this reason, nobody needs to point the others to this person’s blog. Moreover, she tends to write about his experiences and life at an American University, so these posts may not be perceived to be of a general interest outside the community. Finally, *paleofreak* is one of the newest members of this group. Perhaps this blog has had insufficient time to reach enough inlinks. Furthermore, he blogs about ancient animals and science of natural evolution, making the site highly respected and followed, but lacking peers in the community that are able to talk about these topics.

Fig. 9 shows the projections into the map of some outgoing vectors in Table 2. In every case there is a clear movement from the top of the map downwards and towards the end, indicating that blogs with few links start at the top and move downwards as they get older. These “trips” demonstrate that there has been an organization according to age, as linking patterns reflect that age. Old timers tend to be found along the edges of the map, with “founding fathers” found in the lower right-hand corner. The formation of groups is also perceived. For example, it is shown that *rrr* and *javarm* go to the same hexagon in the bottom left-hand corner (Fig. 9). The blogs *atalaya* and *fernando* move towards the bottom right-hand corner. The trajectory followed by each vector along the time is also marked on the map. The labels indicate the vector name and the corresponding period (1 represents the first period, 2 represents the second period and so on).
Analogously to what happens in the map shown previously for incoming links (Fig. 7), blogs with more links are closer to the blog that links them. For example, atalaya, the blog with the most outlinks (and whose set of linked-to blogs is shown in Fig. 10), is closer to fernando (76 links), fbenedetti (29 links), rvr (29) links, www (22 links), blogzine (22 links) ... as is illustrated in Fig. 11.

In general, the SOM provides a global narrative for the community site, and has organized nodes according to their life history. In fact, the SOM organizes sites with a criterion that would be very close to what people following the community would follow: old-timers, active newcomers, and newbies, for instance. In turn, this implies that classification is most related to age than to any kind of standing within the community. Age, as represented by a number of incoming or outgoing links, is the main organizing principle in this community self-organizing map, which places new (few links) sites on one side of the map, and old (many links) on the other side.
Table 3 shows the number of links of the less linked-to blogs along the five periods studied. Fig. 11 shows the blogs which have few links from beginning to end, that is, new or inactive blogs (with 0 or 1 link) stay at the upper zone of the map. Other experiments with few linked-to blogs, not shown in this paper, corroborate that result.

All figures clearly show the trajectory followed by the different blogs along the time, so we could visually predict the future trajectories. Obviously, if we needed an extrapolation more precise we should include more time steps or even we could do maps of specific zones.

Fig. 8. Clustering structure for outgoing links as shown by the U-matrix method: (a) first, (b) second, (c) third, (d) fourth and (e) fifth group of data (outgoing links) projected into the SOM. For this figure, errors were: final quantization error = 0.011 and final topographic error = 0.015.
7. Conclusions

This paper proposes a procedure that visually shows the relationships among different blogs and how these relations are formed along time, considering hyperlinks as indicators of a relationship between members of the same community, and the evolution of blogs from the beginning to its current state.

In order to identify and visualize the emerging communities, the self-organizing maps (SOM) have been used as a powerful tool which produces an unsupervised clustering, and allows the visualization of the projections from a high-dimensional space to a two-dimensional map, highlighting hidden relationships between data set members. The

<table>
<thead>
<tr>
<th>Blog</th>
<th>Number of outgoing links (&gt;100)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Period 1</td>
</tr>
<tr>
<td>atalaya</td>
<td>4</td>
</tr>
<tr>
<td>fernando</td>
<td>23</td>
</tr>
<tr>
<td>pawley</td>
<td>0</td>
</tr>
<tr>
<td>javarm</td>
<td>30</td>
</tr>
<tr>
<td>fbenedetti</td>
<td>0</td>
</tr>
<tr>
<td>rer</td>
<td>50</td>
</tr>
<tr>
<td>jkaranka</td>
<td>0</td>
</tr>
<tr>
<td>jaio-la-espa</td>
<td>0</td>
</tr>
<tr>
<td>eldehwen</td>
<td>0</td>
</tr>
<tr>
<td>verbascum</td>
<td>25</td>
</tr>
</tbody>
</table>
self-organizing maps have been implemented with the help of the SOMToolbox for MATLAB. Via a fast computational procedure, simple operations (such as link extraction), and identification of characteristics of a few sites, the map visually allows to identify communities and what makes them hold together.

All figures help us to analyse the evolution of communities because they show the trajectories followed by the more linked-to blogs and the less linked-to blogs and all of them fulfil clear patterns; the map is clearly divided by age (from top to bottom), number of incoming links (naturally also from top to bottom), topic (left to right, the bottom right zone including technological blogs and metablogs, and the right-hand side formed by more political and personal blogs). In other cases, the organization will not be the same, but still, age, community standing, and topic will be clearly separated in the map.

The working set of websites corresponds to 162 real weblogs (out of a total number of 200; nowadays, there are around 400 blogs, with around one hundred active—posting new stories—any given month) hosted by Blogalia (http://www.blogalia.com/). Only the stories (excluding information in page templates, or dynamic news-feeds, for instance) published in Blogalia were used for this study. There were around eleven thousand in total, which included approximately 17,000 links.

The validity of the proposed method for the identification of communities and for analysing their evolution through time, has been shown empirically using real data. In particular, the trajectories and the expansion of the blogs through the SOM are clearly observed as time passes and as they receive more links. It is clear that blogs form clusters or communities in spite of following different trajectories.

As far as future work is concerned, we will try to develop a methodology that allows further segmentation of the community, and which can be trained online, so that new members can discover others with common interests. We will also update training data, and use time slices, instead of accumulated data, so that a typology of blogs can be extracted by studying this community.
Fig. 10. Projection of the vectors (blogs) that are linked by *atalaya* (in brackets the number of outgoing links to each blog).

Fig. 11. Projections into the map of some vector in Table 3: *fondoazul, elcubo, melicerte and spamzoo*. 
Another future line of work will be to try and predict future position of blogs among the community by comparing the trajectory of a blog to the trajectories of others. This prediction will be more qualitative than quantitative, but we will evaluate what kind of numerical predictions are possible.

Acknowledgements

This paper has been funded in part by project TIC2003-09481-C04, of the Spanish Ministry of Science and Technology, and project PRO39 was awarded by the Quality and Innovation Department of the University of Granada. We are also grateful to Víctor R. Ruiz for his continuous support of the Blogalia community, and to the reviewers of this paper, who have contributed through their comments to its improvement.

References


