THE SPANISH-SPEAKING BLOGOSPHERE: TOWARDS THE POWERLAW?

Fernando Tricas-García
Depto. de Informática e Ingeniería de Sistemas, Univ. Zaragoza
C/ María de Luna, 1
50018 Zaragoza, Spain
ftricas@unizar.es

Juan J. Merelo-Guervós
Depto. Arquitectura y Tecnología de Computadores, Univ. de Granada
C/ Daniel Saucedo Aranda, s/n
18071 Granada, Spain
jmerelo@geneura.ugr.es

ABSTRACT

The blogosphere is the community of bloggers, people or collectives who share information and opinions ordered chronologically. The Spanish-speaking blogosphere contains several thousand blogs; despite its small size, compared to the English-speaking (or maybe global) blogosphere, its characteristics are a bit different.
During the last months the Spanish blogosphere has been growing at a good pace, but we are not sure about whether it has reached its critical mass or not. It is even not clear what would that critical mass be.
This paper will update some numbers provided in previous work, trying to show the evolution during the last months, showing our experience in developing blogging tools, in particular, the "Blogómetro" (http://blogometro.blogalia.com/), which is an open source program that checks on a daily basis the link space in the Spanish-speaking blogosphere, in a similar way to BlogDex or Daypop, which check the English-speaking blogosphere (and a small part of the global one). We will show and analyse data gathered from the end of the year 2002 to November of 2003. The analysis of the data shows that the Spanish blogosphere is approaching a scale-free network behaviour.

KEYWORDS

Weblogs, web-based communities, social networks

1. INTRODUCTION

The Spanish blogosphere is part of the global blogosphere, and is roughly defined as the set of blogs (or, sometimes, blog-looking web pages) that are written in Spanish (in any part of the world) or in any other of the official languages in Spain (Catalan, Basque, Galician). We have also considered blogs written in other languages, if they are written by people living in countries that naturally fit in the Spanish blogosphere area of influence (for example, we heard recently about some Venezuelan bloggers that write in English to point the mainstream attention to their country, or the 'Trilingual blog', http://trilingual.blogspot.com/, or 'kaleboel', http://oreneta.com/baldie/blog/ , written in Catalan, English, and Dutch).

Authors have been observing the Spanish blogosphere since the end of 2002. Our previous work was devoted to present the Spanish blogosphere and some relevant facts about it (Tricas et al. 2003, Merelo et al. 2003). There, some successful initiatives trying to help to construct a sense of community where shown: Secret Cyber-Santa, some channels in 'The internet Topic Exchange' (http://topicexchange.com/), directories, and also some numbers obtained means of our spider and its associated weblog, the "Blogómetro" (http://blogometro.blogalia.com/).

Scale-free networks are interesting because the distribution of connectivity is extremely uneven: some nodes are very well connected, while most of the nodes have only a few links. (Barabasi et al. 1999) discovered that the World Wide Web connectedness follows this pattern. Since in our previous work we

discovered that the Spanish blogosphere did not fit well to this pattern, we wanted to measure if, at least, it could be approaching it. Our hypothesis is that, maybe, the small number of blogs could have influence on this. Perhaps, some minimal number (or critical mass) would be needed in order to reach the adequate situation. In this sense, we try to present here some results about the evolution of this network. Bloggers tend to link other bloggers (and other general sites) so we expect the Spanish blogosphere to show this scale-free network behaviour soon.

The rest of the paper is organized as follows: next section is devoted to explain the framework we have used for this analysis; section 2 will provide some context and information about our work together with some *big numbers*, overall macroscopic measures in the Spanish blogosphere; it will be followed by section 3 which will study the Spanish blogosphere concentrating on its fitting to a power law structure. Finally, we will present our conclusions and some hints on how this work will follow in 4.

2. METHODS, CONTEXT, AND BIG NUMBERS

In this section we are going to provide some explanation about our tools, we will make some comments about the Spanish blogosphere, and we will provide some hints about other projects of interest.

Our project started at the beginning of the Summer of 2002, and it is in a 'beta' stage. All data in this study have been taken from the **Blogómetro**, a suite of tools whose main visible aspect is its blog (http://blogometro.blogalia.com/), hosted in Blogalia (http://www.blogalia.com/). There, a list of fresh links taken from our list of blogs (ranked by the number of sites pointing to them) is published daily. The Blogómetro is an open–source collaborative project, offering our research to the community. It is open to the participation of interested people. In this sense, not only its source code is available at the project page (http://sourceforge.net/projects/blogometro), but also the list of sites scanned daily.

The bot crawls all blogs in the list. From each raw HTML file, it scrapes the links, and stores them in a database if they have not been included before. In consequence, a link is considered *new* or *fresh* if its URL has not been seen before in that particular blog; that means that if a blog refers to another several times by its URL, it will count as a single reference; that also means that links included in the *blogroll* list are considered only once (during the lifetime of the data).

Data has been stored for approximately 12 months, from November, 15, 2002, to November, 1, 2003. We are still collecting data, in order to try to get as much information as possible of the blogosphere, and to be able to study the evolution.

The addition of new blogs to our list is done by hand and, even though we have found moderate interest about the project, we have not received many submissions of sites from other people, so the list is mainly ours (with the pros and cons this may have). Further research will be needed in this area, we hope that people will start sending sites when our project will be more widely known. Comparing with data published in (Tricas et al. 2003) some very successful blog-hosting sites have appeared since then, and that the blogosphere is growing (more than doubled in six months).

During the period of study 723037 links were observed, which yields an average of 2065 links in a day; considering the daily number of blogs, this makes an average of 0.70 links per blog per day. If we assume that each new history posts a new link, we will have around 3500 new histories a day in the Spanish blogosphere. Another measure that can be of interest is the activity of the blogs in our list: during the last month 3060 have posted at least one history, with a link. We can compare these data with the activity in January (1299 blogs), and in November (943 blogs).

3. SPANISH BLOGOSPHERE AND POWER LAWS

The first analysis performed was to check if the Spanish blogosphere fits itself to a power law (Kottke 2003, Shirky 2003). In (Tricas et al. 2003), we fitted a power law $f(x)=kx^a$ using the open-source tool GnuPlot, resulting a=-0.58 and k=328. However, the chi-square test was around 6, much bigger than one, which means that the model does not fit well the data. The blogosphere continues its evolution and now we can say that it still presents a power law structure with a=-0.75, k=2055.34, and a chi-square test of 4.8. It can be seen graphically in Figure 1.

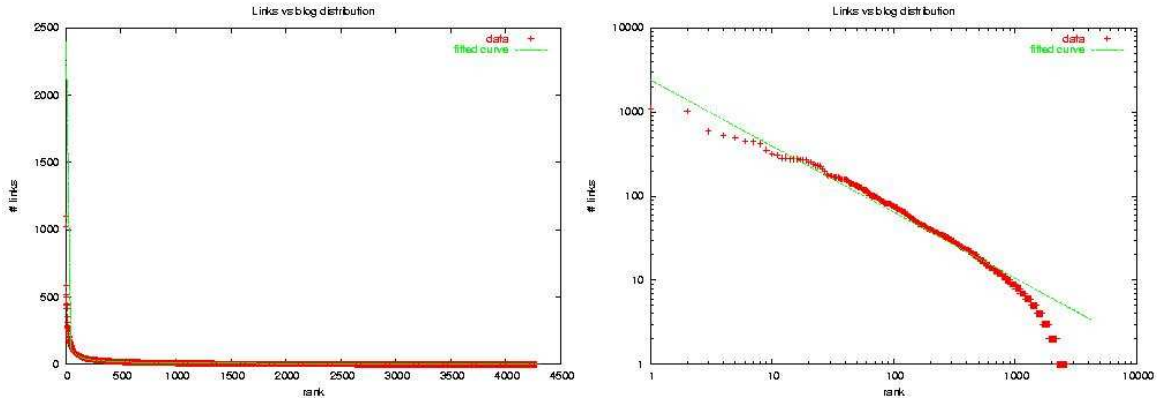

Figure 1. "Power law" distribution of links per blog; x axis represents the blogs ordered by number of incoming links, and y axis the number of links. Data points have been fitted to f(x)=2055.34-0.75; however, the fit is not good enough. The second graphic shows the same distribution on a log-log scale.

Besides the fact that this data does not fit to the model, unlike data published by (Kottke 2003) , where it finds a perfect fit for the (global? English–speaking?) blogosphere, and an exponent of -0.83 (notice that our data are closer now to this exponent).

Our working hypothesis here is that there is some fundamental *critical mass* that makes a certain community behave like a power law; unfortunately (or maybe fortunately), the community we belong to does not seem to have reached that level yet, but, perhaps we are approaching to it.

Our feeling is that, we are approaching to something near to a power law structure, and that as it evolves, we are closer to it. Moreover, some filtering in our raw data could help to get a better fit, since we have been influenced by some spurious links used in different ways by sites (blogrolls, -lists of links to other bloggers-, boxes with other's site contents, ...).

To have a better understanding of these numbers, we can see the monthly variation of the chi-square test adjustment and the graphic showing the evolution of the number of link in Figure 2.

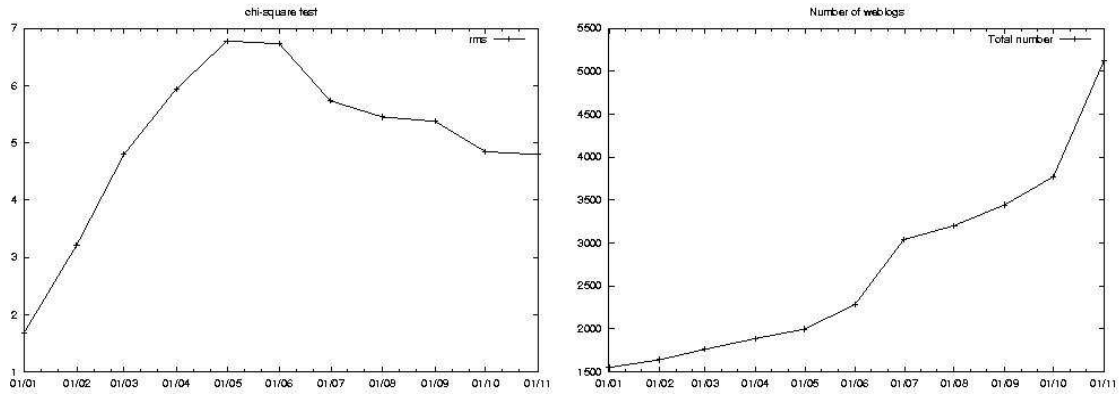

Figure 2. Evolution of the fit to a hypothetic "Power law" distribution of links per blog during this year. Evolution of the number of blogs in our database

4. CONCLUSIONS AND FUTURE WORK

This paper shows the variations in the state of a web-based community, the Spanish speaking blogosphere, during the last year.

The *blogómetro* is an ongoing project, and its measures will be periodically taken to show the state of the Spanish blogosphere, and measure its evolution. We will try to improve the software in several ways,

including public access to the database using web interfaces, addition of self–discovery features, and improvement of other technical details (detecting links that are equal, even if they look different, among others). Until now, we have concentrated on having a tool for studying the blogosphere and helping others to discover it. Maybe we should do more work on the audience aspects, to do the tool known and useful for others. We expect that by in six months or so (maybe more) we will be able to show a power law. We have the feeling that Spanish bloggers do not link so often and frequently as they should and, maybe, we are loosing topics because no links are provided. We are thinking about trying to measure words or phrases, in order to detect topics without links in a similar way to the recently implemented 'Word Burst' of DayPop (http://www.daypop.com/burst/) or 'Memeufacture' (http://memeufacture.com/). A more refined work, separating links to blogs items from links to other general media will be the subject of our research. Another phenomenon that we would like to measure and detect would be what we could call 'background histories'; that is, histories that appear in the blogosphere and are linked slowly during long periods of time accumulating an important number of links, but that do not appear in daily rankings because of this slowness: for sure some of them will be interesting topics for reading.

Other projects we intend to undertake in the future include: cluster formation in the blogosphere (our initial approach to this has been submitted to this conference (Merelo-Guervós et al 2004), interactive visualization, and, if data is available, take measures on other blog communities (such as, for instance, the Portuguese-speaking, which is probably very similar to ours).

Finally, it would be interesting to compare these results with the ones of the global blogosphere, to detect similarities and differences.

ACKNOWWLEDGEMENT

Fernando Tricas-García is with the Group of Discrete Event Systems Engineering (GISED) and his work is partially supported by project CICYT TIC2001-1819, financed by the Spanish Ministerio de Ciencia y Tecnología. Juan Julián Merelo-Guervós work is funded by project TIC2003-09481-C04, of the Spanish Ministry of Science and Technology, and a project awarded the Quality and Innovation Department of the University of Granada. We are also grateful to Víctor R. Ruíz for his creation and continuing support of Blogalia, and his members, for their support during the realization of this work.

REFERENCES

Barabasi, Albert-Laszlo, and Reka Albert, 1999. Emergence of scaling in random networks. *In Science*, 286:509-512, October 15, 1999.

Kottke, J., 2003. Screw the power law, embrace the power law. Available from http://www.kottke.org/03/02/030212screw-power-law.

Merelo et al., 2003. Measuring the Spanish blogosphere. *To appear in Proceedings of Towards New Media Paradigms (COST A20 Conference).*Pamplona, Spain, June 2003.

Merelo-Guervós et al., 2004. Clustering web-based communities using self-organizing maps. *To appear in Proceedings of WBC2004.*Lisbon, Portugal. March 2004.

Shirky, C., 2003. Power laws, weblogs, and inequality. *Available from* http://www.shirky.com/writings/powerlaw_weblog.html

Tricas et al., 2003. Do we live in an small world? Measuring the Spanish-speaking blogosphere. *To appear in Proceedings BlogTalk Conference.* Vienna, Austria. May 2003. *Available from* http://www.blogalia.com/pdf/20030506blogtalk.pdf