

BarriBlog: Virtual neighbourhoods of blogs

M. Luz Congosto

E-mail: mariluz.congosto@gmail.com

Abstract

BarriBlog is a project that aims to examine the content of blogs to determine their identifying characteristics, their social side, the type of contents and their neighbourly ties to other blogs, to study the types of networks they form and they can be visualized.

Introduction

Blogs, one of the tools developed under the "Web 2.0" umbrella, are very popular. They are a personalized means of communication combining text, image and multimedia elements that many people use as an outlet for their creativity.

This simple digital publication tool in log form gave rise to an explosion of contents generated by millions of people. Since 1 April 1997, when **Dave Winer** published what is considered the first post, to date, blogs have evolved and expanded, including a large quantity and great variety of contents. The first blogs, written by professional communicators (proximity to the media) or geeks (technological proximity), were followed by other areas of interest such as politics, feminism, literature, brainteasers, personal reflections, management, advertising, education, science, etc. Very few activities today are not reflected in web logs.

Blogs are descendents of **Internet forums**, which comprise special interest communities that maintain intense conversations, but two features make blogs different: **identity** and **relationships**. On a forum, anyone can begin a discussion. On a blog, the discussion begins with the post located on a site with its own identity, whose author(s) contribute not only contents but also a complete setting: headers, colours, patterns, and other "decorations". This identity appears on the Web in search of relationships. One of the purposes of a blog is "to be heard" and know that someone out there appreciates the effort it takes to maintain a blog.

On the virtual plane, a blog identity is like **a house**, with its doors, windows, furniture, and inhabitants. It is not clear who the "neighbours" are, as nothing brings them together, and yet relations of proximity do exist. There are many different kinds: blogs are recommended by other blogs; contents of other posts are quoted; there are comments that cross among blogs; concordance of contents; etc.

How does one identify a blog's identity? How are the relations of proximity established? How can this be represented visually? These are the questions meant to be answered by **BarriBlog**.

Motivation

I've always felt curious and wanted to see the blogosphere the way you can see our planet from the tool [Google Earth](#). Just as you can see the earth from space and zoom in until you can see the roof on a house, it would be interesting to be able to see the blogosphere as a whole and then focus in on the point you want. That won't be possible for quite some time but we could start building the house from the roof down; that is, take a given blog and attempt to discover its surroundings.

I decided to work on this subject when I read the post by [Julen Iturbe](#) titled **My Neighbourhood: First Life**, which started as follows:

Warning: The characters in this story are real.

I live in a blogosphere that has no name yet. It is a friendly place, a diverse neighbourhood where you can take a walk and strike up a conversation whenever you want. I've lived here for a year and a half. When I moved in, I didn't know any of my neighbours. I know some of them were already living here. And I see lots of others keep coming.

Our neighbourhood is different every morning. Rooms and houses, windows, courtyards, and gardens. It's all rearranged moment by moment. The conversations are what make it so unusual. Each person enters and exits wherever they like. My blogosphere seems a rather odd neighbourhood to those who only skirt the edges but don't actually come in. I tried to find it by using [luistxo](#) in [Tagzania](#), but not even Google Maps was able to locate it.

Current measurements: rankings

Although the blogosphere is supposedly a non-hierarchical space open to conversation and collaboration, rankings were one of the first elements to appear. They separate blogs into castes in a way, reproducing the model of real society. Rankings are generally based on the number of references a blog receives. It is an easy way to measure its importance. However, there are various ranking systems and they do not coincide because they combine the ranking element with other measurements and their scope is different. In the blogosphere in Spain, we have the rankings of [Alianzo](#), [Top Blogs](#) and [blog conversa](#), structured by autonomous communities. In the Spanish-speaking blogosphere, we have [Bitácoras](#) and [Blogalaxia](#), structured by country, and on a worldwide level, there is [Technorati](#).

When one first starts out in the blogosphere, one is guided by rankings, but over time, one tends to stroll through its "long tail" ([larga cola](#)). The way blogs' importance or influence is measured has not evolved much over the years, with the long tail lying outside the measurements that focus exclusively on [A-list bloggers](#).

Overview

The main purpose of BarriBlog is to provide a graphic representation of the proximity relations among blogs. It is based on a set of blogs whose contents are visualized on search engines, internally formed by HTML code, possibly de-structured¹, based on CSS (Cascading Style Sheets). It would be quite complicated and impractical to attempt to find these relations in the content of the blogs, and the optimal results might not be obtained.

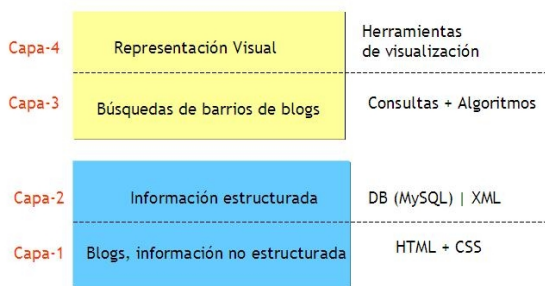
Consequently, information that may be useful must be structured in the search for relations. Currently there is a lot of information about blogs dispersed among various services but some is not freely accessible (Google, Yahoo, etc.) and some is insufficient for analysis (Technorati). In order not to miss information, **one must go to the source**: the contents of the blog.

However, having information about a sole blog is useless; one needs a critical mass large enough to establish relations among the blogs. That requires **on-going discovery of blogs**. Therefore, the extraction of information from a blog should lead to the discovery of new blogs, which in turn lead to other blogs. This recursive task will feed a structured repository of information about blogs, to be used to find the relations among them.

Lastly, results must be obtained from this repository of blogs. The relations among them must be calculated, the type of networks they form must be analyzed, and then a visualization must be made.

This leads us to a four-level model, as shown below:

¹ Los gestores de contenidos de blogs generan XHTML, pero al permitir al autor incluir código HTML generalmente el código se desestructura en un porcentaje grande de blogs.



Level 4 Visual Representation Visualization tools

Level 3 Searches for blog neighbourhoods Consultations and Algorithms

Level 2 Structured information

Level 1 Blogs, unstructured information

A breakdown of the purpose of each level is as follows:

- **Level 1:** Blogs: The original source of data to be analyzed and structured. Contents are HTML (possibly unstructured) based on CSS (Cascading Style Sheets).
- **Level 2:** Structured information about the blog in XML format or in data charts from a relational database.
- **Affinity of contents**
- **Level 3:** Consultation of structured information and search for relations through algorithms.
- **Level 4:** Visual representation of relations.

Conceptual model for BarriBlog

Virutal house

Note: The characterization of the blog is based on the analysis carried out by **Tíscar Lara**. She has characterized 100 personal blogs by journalists for her doctoral thesis.

As mentioned above, in an environment of neighbourhoods of blogs, the basic unit is the “virtual house” each blog represents, as seen below:

Historical record of changes

The house is based on three dimensions:

- **Identity components:** All data identifying the blog, the authors, its host, and the header.
- **Contents:** Number of posts, frequency of publication, most-used tags, comments, quotes from the blog in the posts on other blogs, types of embedded objects, etc.
- **Social side:** All means of external communication or of exploring its social world, including: contact information, syndication, blog-roll, MyTagging, statistics, badgets, widgets, and surveys.

Relations

The relations data gathered correspond to links between blogs, tags, comments, and quotes. To determine the weight of these measurements in establishing the degree of proximity relations, they will be classified as shown below.

Relations are structured along two axes: **Affinity of contents** and **Relation of intensity**. This classification

produces four variations in the relation, which are:

- **Link:** This is the weakest relation between two blogs. It does not imply an affinity of contents but does show a relation between the blogs that may link (0-1) to the other or not. Seen from Blog A, this leads to the following relations: they don't know each other (they are not linked), others know of the blog (incoming link), the blog knows of another (outgoing link), or they know each other (double link).
- **Tags:** This is a measurement of affinity of contents, by coinciding subjects on the blogs. Concordance is measured from 0 to 10 in the top ten tags. 0 is no concordance and 10 means that the 10 most used tags coincide.
- **Conversation:** Comments are a more intense relation. Seen from Blog A, there are four possible states of conversation which can be quantified by intensity: they do not converse (there are no comments between blogs A and B), comments are received (blog B makes comments to blog A at the nth degree), comments are made (Blog A makes comments to B at a degree of n intensity), or both make comments to each other (both blogs exchange comments). The quantification is the result of "total number of comments / total number of posts".
- **Quotes:** This is the most intense relation among blogs, given that it denotes affinity in contents as well as an intense degree of relation. Seen from Blog A, the following relations are observed and can be quantified by intensity: they do not quote each other (there is no trackback between the blogs), the blog is quoted (blog B quotes blog A at the nth degree), the blog quotes (blog A quotes blog B at the nth degree), or they quote each other (they interchange trackback).

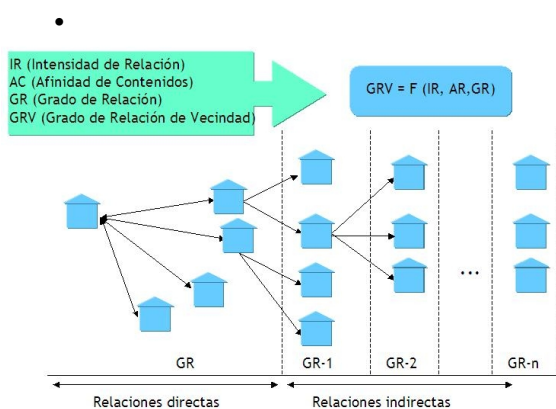
Evaluation of the degree of proximity

The previous classification gives us three parameters:

- **Intensity of relation (IR)**
- **Affinity of contents (AC)**
- **Degree of relation (GR):** Measures the distance of links between two blogs. It is a variable parameter that is introduced once a blog's neighbourhood of blogs is discovered. Its value can range from 5-1. This measurement is used to set the limits of a blog. To determine a "first degree" neighbourhood of blogs, only those that have direct links to each other are included. For the "second degree", blogs that are directly related and also those indirectly related at the second level are included.

The **degree of proximity** GRV will be determined from these values, which will be a function of the parameters IR, AC and GR

The figure below represents the relation scenario



Direct relations Indirect relations

Visualization

The huge amount of information will be condensed, or distilled, to obtain the magic amount of the degree of proximity (GRV) enabling the visual representation of the neighbourhoods of blogs, from the traditional graph method to representations more in keeping with the “urban development of blogs”.

Los gestores de contenidos de blogs generan XHTML, pero al permitir al autor incluir código HTML generalmente el código se desestructura en un porcentaje grande de blogs.