Studying Blog Features over Link Popularity

José Luís Devezas[‡]
joseluisdevezas@gmail.com

Cristina Ribeiro[†‡]
mcr@fe.up.pt

Sérgio Nunes[‡]
ssn@fe.up.pt

[†]INESC-Porto
[‡]DEI, Faculdade de Engenharia, Universidade do Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto, Portugal

ABSTRACT

The study of the blogosphere can provide sociologically relevant data. We analyze the links between blogs in the portuguese blogosphere, in order to understand how they group and interact, to identify clusters and to characterize them. Our data set contains post data for more than 70,000 blogs, with over 400,000 links. The linkage data is represented as a blog graph and partitioned into several slices, according to their in-degree. We then study the evolution of blog features, and observe a consistent pattern of decrease in posting frequency, number of out-links, and post length, as we move from the highly-cited blogs to the less cited ones.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Clustering, Information filtering, Selection process

General Terms

Measurement; Experimentation

Keywords

Blogosphere characterization, Link analysis, Blog clustering

1. INTRODUCTION

With the prominence and magnitude of the World Wide Web comes the opportunity to study the behaviors of online communities. We analyze the blogosphere as one of these online communities, combining link analysis with blog characterization. While blogs have specific features and behaviors, they also interact with each other, interconnecting by means of hyperlinks, in the World Wide Web. When studying the blogosphere, a lot of information can be extracted from the link structure, namely the most cited blogs, the central blogs and the densely connected groups of blogs that may reveal communities. We analyze the links between blogs in the portuguese blogosphere in order to understand how blogs group and interact, to identify clusters and to characterize them. We use a representative data set of the portuguese blogosphere previously studied by Couto [3] to analyze the characteristics of several blog clusters that share a similar popularity rank. For the analysis, we partition the blog graph into several slices, ordered by decreasing in-degree, and study the evolution of features for progressively less cited slices. Observed features include post creation over time and number of links and words per post. We identify the most cited blogs and study the differences in behavior between the highly cited and the least cited blogs and the characteristics of each of these sets.

2. BLOGOSPHERE ANALYSIS

Network analysis has already been applied to the web and also to the blogosphere. Broder et al. [2] studied web connectivity, using a collection of over 200 million pages and 1.5 billion links. They observed that the web divides into three main components — a strongly connected component (SCC), a set of pages connecting to the SCC and a set of pages the SCC connects to — in what has become known as the "bow-tie" model. Kumar et al. [7] introduced the concept of time graph and used prefix graphs to characterize the evolution of the blogosphere, analyzing degree distributions, the size of the strongly connected components, the number and size of communities and the bursts of activity inside the communities. Our work is also supported on a blog graph with several vertex and edge properties. We focus on comparing the behaviors of blogs depending on their popularity. Blog characterization is also a topic that has already been explored. Similarly to Herring et al. [6], we analyze several blog features, like the number of words and the number of out-links. The novel factor that we introduce with our work is the blog characterization over link popularity as opposed to, but not disregarding, the analysis over time. We aim at distinguishing popular blogs from the remaining by studying the behaviors of several blog clusters that share a similar in-degree.

3. THE BLOG COLLECTION

The collection we use for our analysis was provided by SAPO, a portuguese Internet service provider and owner of SAPO Blogs [8], a popular portuguese blog hosting service. This collection is made of a set of posts, written in portuguese, from various blogging services, mostly SAPO Blogs and Blogger. Since we are unaware of the criteria used to select portuguese blogs outside of SAPO's domain

1st Workshop on Social Media Analytics (SOMA '10), July 25, 2010, Washington, DC, USA.

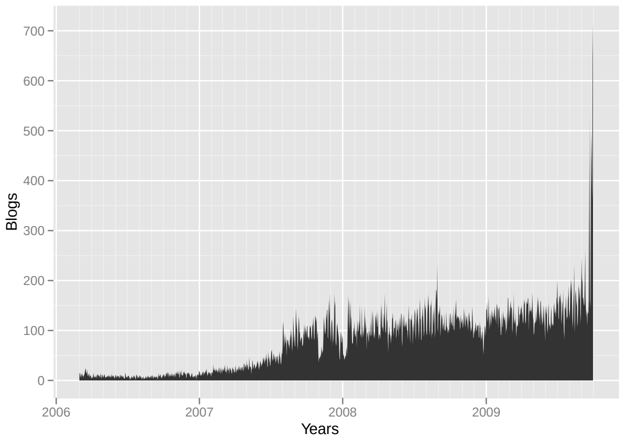

Figure 1: Newly created blogs per day.

and cannot ensure the thoroughness of that data, we decide to focus our study on SAPO's blogs, which have previously been determined as representative of the portuguese blogosphere [3]. The data set is not a raw collection harvested by SAPO, we are aware that the ISP also performs some cleaning on the collected data; we do not know the criteria used for excluding blogs and assume it is just a very basic splog exclusion. The collection compiles more than 100,000 blogs, with over 2 million posts, from which we build a data set with over 70,000 blogs and 400,000 links, extracted from a table with approximately 17 GB. Each blog is hosted under the "blogs.sapo.pt" domain using a user-defined subdomain, in the format "blogname.blogs.sapo.pt". We have access to posts with dates ranging from March 1st 2006 to October 1st 2009. This means that the data set compiles post data starting from the latest release of SAPO Blogs until the month we start our analysis. We do not, however, consider October 2009, because we only have posts for the first few days of that month. All the data is stored in a MySQL relational database management system and, for each post, we have access to a series of fields, from which we only use the ID, the URL, the creation date and the actual content of the post.

3.1 Data Set Validation

Prior to extracting and arranging the data for network analysis, we do a standard analysis on the blog set. We depict the newly created posts and blogs, per day, over the years. The results prove to be less straightforward than we expected. Figure 1 shows the number of new blogs, created per day, over the years. During the last month — September 2009 — there is a peak that stands out. We determine that the average value for newly created blogs per day is 79.09 and the median value is 88. Having a value of 715 for the newly created blogs on the last day of September 2009 appears to be an irregularity. In an attempt to understand and explain this spike, we manually browse through some of the blogs created on September 30th 2009 and verify that many don't exist anymore. Since this verification is being made less than one month later, it all indicates that those blogs have been deleted by SAPO for being spam blogs (splogs). We create a script that, for a given day or month, returns a list of blogs that don't exist anymore — this means the HTTP request either returns a "404 Not Found" error or a SAPO web page with the information that the user is unknown. We run the script for September 2009 and get a list of 3,187 bogus blogs — 42% of September 2009 blogs don't exist on October 21st 2009. Even though some of the blogs might have been deleted by their owners, the percentage of bogus blogs represents almost half of the blogs for that month and most of the usernames for those blogs seem to be computer generated, with very few exceptions. By running the script for the rest of the months of 2009 — January to August — we verify that, in average, 22% of the blogs don't presently exist on the web. Though September 2009 is the most recent month of the collection, it already has twice as much bogus blogs than the average for the previous months of the same year. We speculate that SAPO frequently removes splogs from the collection and that the spike on the chart is associated with a time frame when the cleaning process hasn't yet been applied. Being unsure of how representative that month really is and given that it is out of our scope to study spam blogs, their detection and removal, we decide to simply leave September 2009 out of our analysis.

4. LINK ANALYSIS

In order to analyze the link structure of our sample we represent the blog link structure as a graph, associating several attributes with the vertices and edges. Each vertex represents data about a blog, and each edge represents data about a link — including data about the post where the link was extracted from. We propose that a general approach to study blog features over link popularity should iterate through the following steps until results are satisfactory:

1. Select and apply a criterion for blog clustering based on the link structure.
2. Identify and analyze several features for the clusters.
3. Distinguish or categorize the clusters based on their behaviors.

For our case study, we are interested in understanding whether popular blogs behave differently from less popular ones. In a first approach to step 1, we generate blog clusters by partitioning the blog graph into several slices ordered by the number of citations. Our criterion is that blogs with similar in-degree stay in the same cluster. Given the dimension of the collection, we adopt a size of 1,000 for the sets of blogs. The first slice contains the 1,000 blogs with the highest in-degree. The features we study for each slice are post creation over time, number of out-links per post and number of words per post. Analyzing these features depending on the popularity rank gives us an overview of how the behavior evolves when we go from the most popular to least popular blogs.

4.1 Data Preparation

Our analysis requires accessing the collection and extracting the links we will be working with. Data preparation involves querying the MySQL database, filtering the records by hostname and date interval, parsing the post bodies and indexing the extracted linkage data using a Berkeley DB key/value database. The result is data aggregated by hostname, disregarding links that point to a host outside the

Table 1: Information stored in the blog graph.

	Attribute	Description
Blogs	`name`	Blog hostname.
	`date`	Creation date of the oldest post.
Posts	`post.url`	Complete URL for the link source
	`post.date`	Creation date of the post.
	`post.wordcount`	No. of words of the posts's content.
	`post.charcount`	No. of characters of the post's content.
Links	`name`	Complete URL for the link target.
	`source`	Link source blog node.
	`target`	Link target blog node.

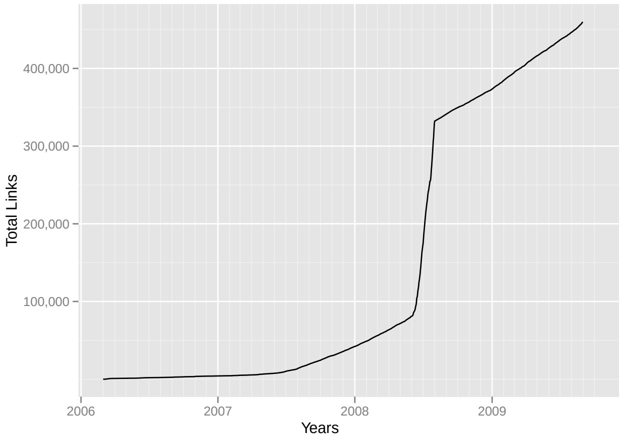

Figure 2: Cumulative number of out-links over the years.

collection and any invalid hostnames that may have been extracted from malformed HTML. Based on the resulting nodes, we go through the index and generate a GraphML [1] representation of the blog graph — possibly with edge multiplicity and loops — that is loaded into the igraph [4] network analysis tool. The resulting blog graph is made of 72,591 vertices and 459,737 edges. Table 1 summarizes the attributes considered in the graph structure — blog attributes are associated with the vertices and post and link attributes are associated with the edges.

4.2 Link Usage

Prior to the cluster analysis, we depict the evolution of the total number of links over time (Figure 2). In average, the link collection of the blog sample grows 17.88% per month. There is a link creation activity peak during the months of June and July 2008, which seems to indicate that link usage has become more prominent one year after the burst in the blog and post creation activities.

4.3 Slice Characterization

Ordering the blogs by decreasing in-degree, we partition the blog graph into slices of 1,000 blogs each, keeping link multiplicity and self-citations, and compute the mean and median values for blog features in each slice. So, for example, when we look at Figure 3, the slice of order 0 represents the group of blogs ranked from 1 to 1,000, and in the y-axis we find the mean and median values for the monthly posting activity of this group of blogs (7934 and 3637, respectively); the slice of order 10 represents the group of blogs ranked from 10,001 to 11,000, having the values of 44 and 43 for the mean and median, and so forth. Table 2 shows the mean and median in-degree and out-degree for the blogs in each slice. As expected, per definition, the slices with lower order have the highest in-degree. The out-degree also follows a similar distribution. For the blogs in slice 0, we have an average in-degree of 371 and a median of 71. The difference between these two values is evident. It means that, even though there are some blogs with a rather high in-degree, most blogs in this slice have a less than average number of in-links. From slice 1 onward the mean and median values are a lot more similar to each other. From slice 12 to slice 19 the in-degree is already as low as the unit and from slice 20 onward the in-degree is null. When analyzing the out-degree, the mean and median values are also decreasing, with a null median for all slices following slice 9. On the other hand, the average out-degree is never null, even for the slices with the least popular blogs — from slice 17 onward there are between 100 and 1,000 links per slice. We verify that both the values for newly created posts per month (Figure 3) and monthly number of out-links per post (Figure 4) are higher for blogs with a high in-degree, suggesting that blogs that frequently generate new content and link to other blogs tend to be the more cited. Regarding the evolution of the number of words per post (Figure 5), we verify that this value decreases progressively, as blogs become less cited, showing a clear relation between post length and popularity — blogs with more content tend to be more cited.

Table 2: Slices in-degree and out-degree.

Slice	In-Degree		Out-Degree	
	Mean	Median	Mean	Median
0	371.0	71	333.2	62
1	26.6	26	24.9	21
2	15.1	15	14.3	12
3	10.1	10	10.6	7
4	7.3	7	6.8	4
5	5.5	6	5.6	2
6	4.3	4	4.4	2
7	3.4	3	3.5	1
8	2.9	3	2.7	1
9	2.0	2	1.9	1
...	...	...	...	...

5. CONCLUSIONS

We have studied a large sample of the portuguese blogosphere, by partitioning it, using the number of citations as criteria, and have examined several behaviors for those parts, in order to understand whether the most popular blogs have distinguishing characteristics. By analyzing the mean and median values for the monthly newly created posts and the number of links and words per post, we identified a consistent pattern of change as we move from the highly cited blogs to the less cited. The characteristics of the slices in what concerns posting activity, linking pattern and number of words varies strongly. There are evident differences between the highly cited slices and the remaining ones, illustrating the contrast between popular and less popular blogs.

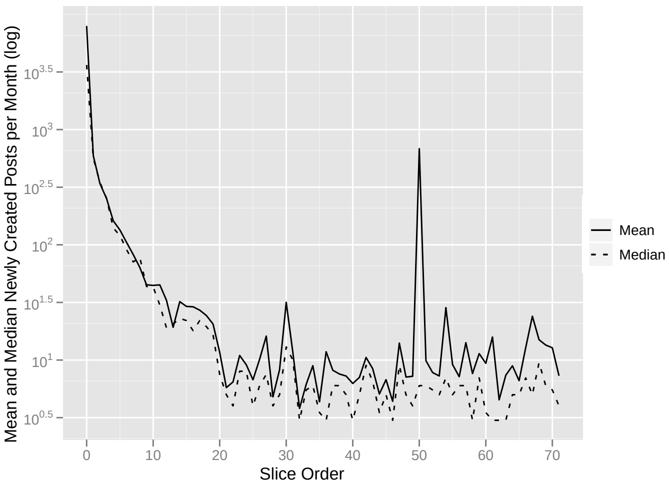

Figure 3: Newly created posts per month.

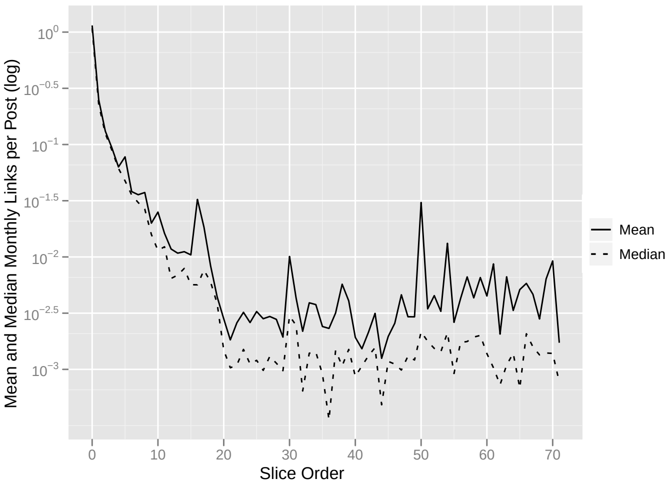

Figure 4: Monthly number of out-links per post.

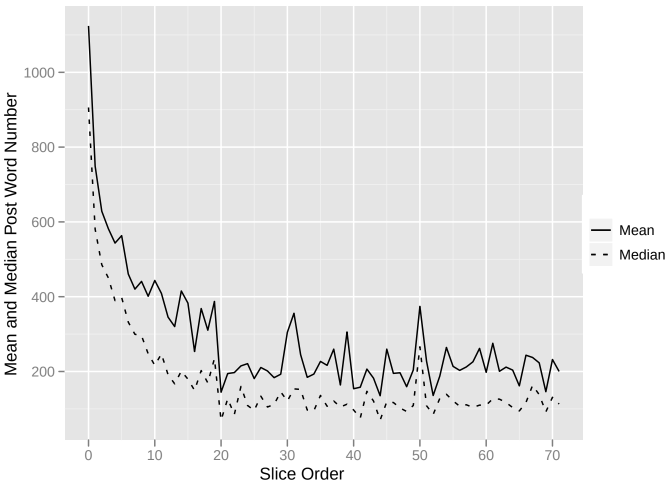

Figure 5: Number of words per post.

6. FUTURE WORK

The analysis of the portuguese blogosphere's link structure leaves an open door for future studies. Based on the methods used by Kumar et al. [7] for the analysis of the bursty evolution of the blogosphere, we could study the evolution of blog popularity, in order to understand what influences a blog to become a reference in the blogosphere. Several prefix graphs, for different time frames, could be extracted from the blog graph and partitioned using a ranking heuristic. Having different blog members for the corresponding slices of each prefix graph, an analysis of the rank evolution of the most popular blogs could be made, accompanied by a study of their features evolution. Another lead for future research is the detection and characterization of portuguese blog communities, eventually applying the algorithms implemented in igraph to detect densely connected subgraphs. We could also study the link polarity for the various communities, identifying whether a community is densely connected because it loves a certain subject or because it negatively criticizes it, and perhaps identify a group of central blogs as the target of discussion. Finally, based on our previous work in this area [5], we could research spam detection processes, apply them to the blog graph and do an analysis of the features over link popularity on the resulting structure, possibly comparing the outcome to the unclean version of the graph.

7. ACKNOWLEDGMENTS

We would like to thank SAPO, a portuguese ISP and blog hosting service, for providing the data set that served as base to the study and Telmo Couto for providing information about his previously developed work on the characterization of an earlier snapshot of the same collection.

8. REFERENCES

[1] U. Brandes, M. Eiglsperger, and J. Lerner. GraphML Primer. In *Graph Drawing*, 2005.

[2] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the Web. *Computer Networks*, 33(1-6):309 – 320, 2000.

[3] T. Couto, C. Ribeiro, and S. Nunes. Characterizing the Portuguese Blogosphere. In *International Conference on Weblogs and Social Media (ICWSM'09)*. AAAI, 2009.

[4] G. Csárdi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[5] J. L. Devezas. Link Ecosystem of the Portuguese Blogosphere. Master's thesis, Faculty of Engineering of the University of Porto, 2010.

[6] S. Herring, L. Scheidt, I. Kouper, and E. Wright. A Longitudinal Content Analysis of Weblogs: 2003-2004. 2007.

[7] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the Bursty Evolution of Blogspace. *World Wide Web*, 8(2):159–178, 2005.

[8] SAPO. Blogs do SAPO. http://blogs.sapo.pt, October 2009.