Characterizing the Portuguese Blogosphere

Telmo Couto, Cristina Ribeiro, Sérgio Nunes
Departamento de Engenharia Informática
Faculdade de Engenharia da Universidade do Porto
Rua Dr. Roberto Frias, s/n 4200-465 Porto, Portugal
telmocouto@gmail.com, {mcr,ssn}@fe.up.pt

Abstract

Over the years, blogs have become a popular and influent part of the web, having many distinctive features — they are usually thematic and organized by communities, and their contents have a chronological order. In this paper, we present a characterization of the portuguese blogosphere based on a large set of portuguese blogs collected from multiple sources across a 5-year period. The extent of our collection allows us to draw some conclusions on the evolution of the blogging activity over the years. We survey and apply methods that have been used for web and blog characterization and, using samples of blogs obtained from other sources, we try to estimate the extent to which this collection can be representative of the portuguese blogosphere. The paper presents multiple statistics that allow us to characterize the portuguese blogosphere and compare it to results obtained with other studies.

Introduction

The blogging activity has become a popular and important form of communication on the web over the years, with more than 133 million blogs indexed by Technorati as of August 2008 (Sifry 2008). Typically, they consist of special web pages with chronologically ordered entries, often with embedded links and occasionally a few images, and they are usually thematic. Blog entries (or posts) include a link that refers to themselves so that others can refer to individual posts. While a typical Web page has a single URL as its point of entry, blogs have multiple locations of interest, since posts have individual value (Cohen & Krishnamurthy 2006). In August 2008, the worldwide blogosphere was estimated to contain over 133 million blogs, from which an estimated 2% are written in Portuguese (Sifry 2008).

We define the portuguese blogosphere as the set of blogs in the global blogosphere that are written in Portuguese by people living in Portugal or by portuguese people living in foreign countries. Although the first portuguese blogs have been around since the 90's, the portuguese community in general wasn't yet very familiar with blogging activity by the year of 2006 (Cheta 2008). It was in this year that SAPO, a major portuguese ISP, launched a blogging service dedicated to the portuguese community.

As a result from a collaboration with SAPO, we were handed a large collection of portuguese blogs that spans across a long period of time and contains blogs from multiple sources. Based on this collection, we aim to provide a characterization of the portuguese blogosphere and how it evolved over the years. Besides, the extent of the collection allows us to draw some conclusions on the overall blogging behavior among all bloggers using a specific service that is characteristic of a national community.

Related Work

Early studies on the blogosphere discussed the differences between blogging and journalism and the influence of external events on the blogging activity, stating that most blogs were used as personal journals and observing that links were less frequent than on the web (Herring *et al.* 2006). Studies of reference include the creation of the TREC Blogs06 Collection (Macdonald & Ounis 2006) and the annual "State of the Blogosphere" reports (Sifry 2008), which provide statistics for multiple characteristics of the blogosphere. The usage of hyperlinks in blogs and the detection of trends and authorities has also been the subject of many studies focused on the detection of trends and authorities in the blogosphere (Cohen & Krishnamurthy 2006). Research on the persian blogosphere also took advantage of a large dataset of blogs containing comments (Qazvinian *et al.* 2007).

Research on the portuguese web found that a large amount of websites had only one document and that 93% had less than 100 documents. Among all the presented statistics, an interesting observation was that most of the portuguese web pages didn't have any links to other portuguese sites (Gomes & Silva 2005). A survey of a sample of the portuguese population in early 2006 found that only a third of the respondents considered themselves as internet users, but 55% of those claimed knowing what a blog was. A quarter of the internet users had a frequent practice of browsing through blogs, usually interacting with them through comments and e-mail, and half of those had their own blog. In general, the blogging activity hadn't been adopted by most of the portuguese internauts at that time (Cheta 2008).

Most of these studies were based on collections containing between 20,000 and 50,000 blogs. Considering this aspect, our collection stands above the average with more than 60,000 blogs in corpus. Previous studies in this blog collection detected an irregular growth in the number of blogs over time, while observing that the portuguese bloggers link more often to themselves than to other blogs, with the number of links in posts increasing over time (Branco 2008; Pinto 2008).

While some characterizations were focused in deep studies of a few characteristics, others provide a more generic analysis of multiple aspects of the blogging activity. An extended version of the work described in this document is presented in (Couto 2009).

The SAPO Blogs Collection

Although the first portuguese blogs were created in the late 90's, it was only in recent years that they have gained more expression. In March 1st, 2006, SAPO, a major portuguese ISP, launched SAPO Blogs[1], a blogging service dedicated to the portuguese community. The collection we use in our research was created in order to provide a convenient blog dataset for scholar research on the portuguese blogosphere, and is based on a dump from SAPO's database, containing over 60,000 blogs and 3,5 million posts. It differs from other blog collections, in that it contains the entire set of blogs hosted in the SAPO Blogs service and a set of blogs from other providers collected through a crawling process over the web.

Representativeness of the Collection

One of the main advantages of this collection is that it covers a vast period of time, containing blogs and posts created until the end of June 2008, which allows for research on the evolution of the activity in the portuguese blogosphere over the years. However, it must be taken into account that, since a part of this collection was built by a crawler that worked over the links found inside blogs, an important part of the portuguese blogosphere might have been left out. Hence, it is important to estimate the extent to which the collection can be representative of the portuguese community.

Since there are no simple measures that can be applied to evaluate the representativeness of a collection, we decided to focus our analysis on the number of portuguese blogs covered by our collection and the distribution of blogs by service provider. As a reference, we used Google's Blog Search[2] engine to compare search results with our collection's contents. The main reasons behind this choice are Google's reputation as an effective service, believing that its results are representative of the blogosphere, and the tools it provides that allow an easy automation of the process.

Blog Coverage

The method used to estimate the blog coverage within the collection was to query Google Blog Search for portuguese blogs that would contain specific terms and subjects considered representative of the portuguese reality, extract the results and compare them to our blog collection. With Google's tool, we retrieved the links to blog posts that contained the queried terms, used the posts to identify unique blogs and calculated the number of those blogs that were also found in the collection. We defined the Blog Ratio as the percentage of blogs matching our collection within all blogs found with Blog Search.

Provider	Collection	Google Blog Search
SAPO	64.8%	12.9%
Blogger	33.9%	71.2%
Other	1.3%	15.9%

Table 1: Distribution of blogs by provider in different sets.

This method has some limitations. First, we are assuming that the results provided by Google Blog Search are reliable and representative of the blogosphere for any given query. The second biggest limitation is related to the queries to use. Since there is no possible way to filter only portuguese blogs from the Blog Search results, our queries must be associated to the portuguese culture and national events that are less likely to be commented in other portuguese-speaking countries, like Brazil.

Selecting the queries to use was the biggest challenge for the retrieval of the Blog Ratios. Pinto's work on the detection of blog trends and popular topics (Pinto 2008) was an important step in the study of blog contents within the portuguese blogosphere, so we decided to use some of the most popular topics detected in his work as a base for our queries.

The results obtained with these queries were satisfactory, with an average 51.6% of the blogs retrieved being found within the collection. In order to validate these results, we tested a new set of queries related to more recent events and some informal linguistic expressions. The new results were very similar to the previously obtained, validating our claim that half of the portuguese blogs found by Google's Blog Search engine can also be found within our collection.

Blog Service Providers

Our collection was built upon all the blogs hosted by SAPO, followed by a crawl that tracked links from these blogs to find portuguese blogs hosted in other servers. Since the collection's contents were expected to be biased towards the blogs hosted by SAPO, we decided to inspect the expression of the different blog service providers within the collection.

During our analysis, we came across 82,961 blogs without posts in the collection, with 97% of them being hosted by SAPO. We inspected a sample of these empty blogs from SAPO in the web and found that most of them actually existed as empty blogs. Unlike other providers, the SAPO Blogs service allows people to register their blog domains without requiring an introduction post. Therefore, most of the empty blogs within this collection belong to people who probably intended to start a blog and even registered a domain, but never wrote a post in that blog. This accounts for nearly 67% of all blog domains registered in the SAPO Blogs service. However, the empty blogs provide very little information, so we had to filter them out for this research.

[1] http://blogs.sapo.pt

[2] http://blogsearch.google.com/

Provider	Blogs	Posts	Average
SAPO	40,045	818,797	20.45
Blogger	20,943	2,630,178	125.59
Other	842	113,651	134.98
Total	61,830	3,562,626	57.62

Table 2: Number of posts per blog in collection, by provider.

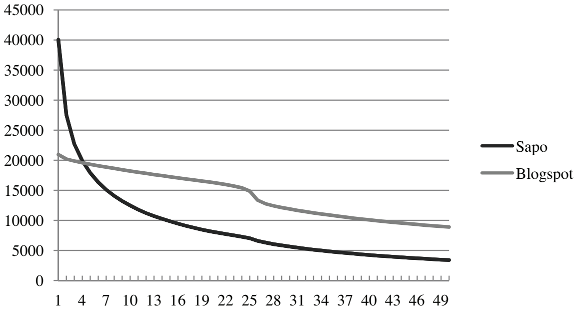

Figure 1: Number of blogs with n or more posts, by provider.

We retrieved the number of blogs from each provider in the collection and compared them to the sample of blogs previously retrieved with Google Blog Search. The distribution of blogs by provider, as can be seen in Table 1, is very different between the two sets — 65% of blogs from the collection are hosted by SAPO, while 71% of blogs from Google Blog Search are hosted by Blogger[3]. These results seem to be inconsistent with the results from our blog coverage analysis: if there is such a big difference between the blogs from our collection and those found in Google Blog Search queries, how could our Blog Ratios have values near 50%?

To analyze this situation, we decided to calculate the average number of posts per blog in the collection according to the service provider, as shown in Table 2. We found that, while the collection contains an average of 57.6 posts per blog, there's also a large difference between the average number of posts from blogs hosted by SAPO and all the other blogs.

We retrieved the number of blogs hosted by SAPO and Blogger that reached a specified number of posts, as presented in Figure 1. We observed that half of the blogs hosted by SAPO have ended at 4 posts or less, while only a few blogs from Blogger have ended below that limit. This can help to explain the inconsistency above. Blogs with fewer posts are less likely to be found, and so the subset of blogs from Blogger in our collection tends not to include them. Blogger's curve drops suddenly at the number of 25 posts. We believe that this is due to a problem with the feeds used; the crawling process was unable to retrieve more than the 25 most recent posts from those feeds, which might be caused by the settings of the respective blogs. As the number of posts increases, the curves from both providers have increasingly similar behaviors — which might indicate that, if we had the entire collection from other providers, the curves would be similar to SAPO's from the beginning.

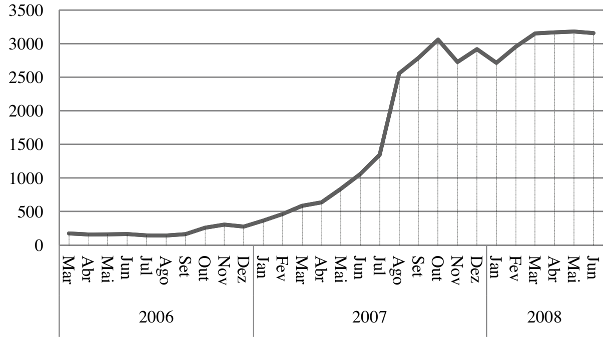

Figure 2: Number of new blogs created per month.

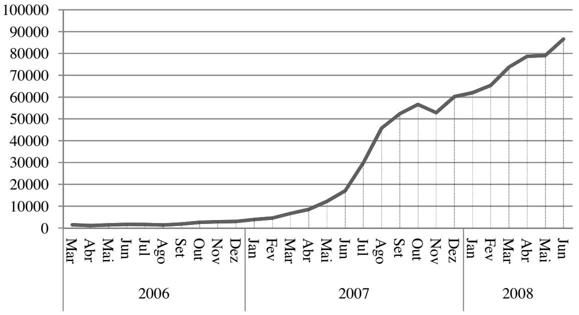

Figure 3: Number of new posts created through time.

We also analyzed the distribution of posts by the time of day and observed that the posting activity is very similar between the two groups. In general, our results indicate that blogging behaviors within the community are independent from the blogging service used.

Characteristics of the Portuguese Blogosphere

We observed that the collection presents a good coverage of the portuguese blogosphere. Although the blogs within the collection aren't representative of the reality in terms of blog service usage, we also observed that the blogs hosted by SAPO form a complete dataset from a portuguese blog service provider that sport a similar behavior to other known portuguese blogs. For these reasons, we considered this subset from the collection to be representative of the portuguese blogosphere in a smaller scale and decided to focus our characterization on the set of blogs hosted by SAPO.

The growth of the blogosphere is often evaluated by the growth in the numbers of blogs and posts. Figure 2 depicts the number of new blogs created per month since the launch of the SAPO Blogs service. During the first year of activity, the number of new blogs created had a very slow growth. It was during the year of 2007 that activity in the service

[3] http://www.blogspot.com

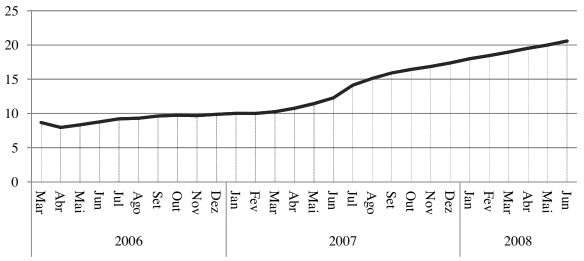

Figure 4: Average number of posts per blog through time.

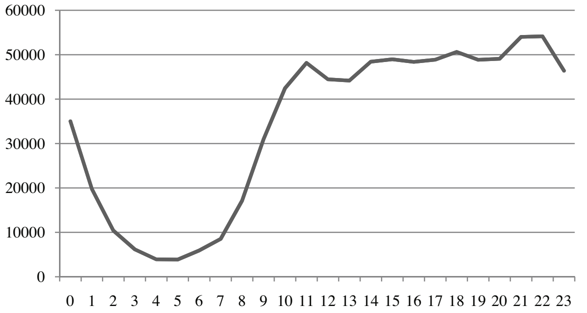

Figure 5: Post distribution per hour.

really took off, with more than 2,500 blogs being created per month. This trend seems to follow the same growth pattern observed in the worldwide blogosphere (Sifry 2007).

The number of new posts created per month over time can be observed in Figure 3. During the first year of activity, the posting behavior is similar to the one presented by the creation of new blogs. However, after July 2007, the posting activity started increasing at a much higher rate. This observation led us to retrieve the cumulative numbers of blogs and posts created over time, in order to estimate the growth of the posting activity.

Figure 4 presents the evolution of the number of posts per blog over time. During the year of 2006 and the first half of 2007, blogs had an average of 10 posts but, since then, this average increased significantly. As of June 2008, the number of posts per blog had nearly doubled with respect to the same period in the year before. Despite the high number of bloggers who write a few posts but don't follow through, we observed that the posting activity has increased over time, indicating a solid growth in the number of bloggers that create new content in the blogosphere.

The distribution of posts according to the reported hour of posting can be observed in Figure 5. The portuguese bloggers usually post less during the morning. Activity peaks between 11 and 12 o'clock and then slows a bit during lunch time, which remains a high activity period though. Most of the posts are submitted during the working schedule after lunch, but there is also a clear peak of activity after dinner.

Conclusions

We observed that the portuguese blogosphere has increased over the years, in the numbers of new blogs and posts created. In June 2008, the average number of posts in portuguese blogs was 20, nearly doubling in relation to same period in the previous year.

Current work is now focusing on the analysis of the link structure within the portuguese blogosphere and how it evolved over the years. Identifying the relationships between blogs within the blogosphere and outside of it and how they relate with other types of media is expected to provide insight into the detection of authorities among portuguese bloggers.

Acknowledgments

We would like to thank SAPO for making the blog collection used in this work available for our research.

References

Branco, J. M. 2008. Aplicação do H-Index em Blogues. Master's thesis, Faculdade de Engenharia da Universidade do Porto.

Cheta, R. 2008. Bloguers e Blogosfera .pt. Technical report, OberCom.

Cohen, E., and Krishnamurthy, B. 2006. A Short Walk in the Blogistan. *Computer Networks* 50(5):615–630.

Couto, T. 2009. Characterizing the Portuguese Blogosphere. Master's thesis, Faculdade de Engenharia da Universidade do Porto.

Gomes, D., and Silva, M. 2005. Characterizing a National Community Web. *ACM Transactions on Internet Technology (TOIT)* 5(3):508–531.

Herring, S.; Scheidt, L.; Kouper, I.; and Wright, E. 2006. A Longitudinal Content Analysis of Weblogs: 2003-2004. *Blogging, Citizenship, and the Future of Media* 3–20.

Macdonald, C., and Ounis, I. 2006. The TREC Blogs06 Collection: Creating and Analysing a Blog Test Collection. *Department of Computer Science, University of Glasgow Tech Report TR-2006-224*.

Pinto, J. P. 2008. Detection Methods for Blog Trends. Master's thesis, Faculdade de Engenharia da Universidade do Porto.

Qazvinian, V.; Rasoulian, A.; Shafiei, M.; and Adibi, J. 2007. A Large-Scale Study on Persian Weblogs. In *Proc. of Workshop on Text-Mining and Link-Analysis*.

Sifry, D. 2007. Sifry's Alerts: The State of the Live Web, April 2007.

Sifry, D. 2008. Technorati: State of the Blogosphere 2008.