FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

FEUP

Characterizing the Portuguese Blogosphere

Orlando Telmo de Oliveira Gomes e Couto

Report of Dissertation
Master in Informatics and Computing Engineering

Supervisor: Cristina Ribeiro (Aux. Professor)

Co-Supervisor: Sérgio Nunes (Lecturer)

3^{rd} March, 2009

Characterizing the Portuguese Blogosphere

Orlando Telmo de Oliveira Gomes e Couto

Report of Dissertation
Master in Informatics and Computing Engineering

Approved in oral examination by the committee:

Chair: João Carlos Pascoal de Faria (Aux. Professor, FEUP)

External Examiner: Pável Pereira Calado (Aux. Professor, Instituto Superior Técnico)
Internal Examiner: Maria Cristina de Carvalho Alves Ribeiro (Aux. Professor, FEUP)

19^{th} March, 2009

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, their contents have a chronological order and they are interesting both for social studies and for diverse economical activities. In this dissertation, 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.

We survey and apply methods that have been used for web and blog characterization. Using samples of blogs obtained from different sources, we observe that a satisfactory ratio of portuguese blogs is covered by the collection. We analyze the characteristics of this collection and study the nature of the blogs from different providers in the dataset, in order to identify subsets with different properties that can be valuable for research. We examine the different subsets in multiple behaviors, including the evolution of the corpus over time and the number of posts per blog, and determine the extent to which this collection can be representative of the portuguese blogosphere.

Our characterization of the portuguese blogosphere is based on a set of blogs from this collection that contains the entire dataset from a portuguese blog service provider. The extent of this particular set allows us not only to draw some conclusions on the blogging activity over the years, but also to study the blogging behavior among users of a particular blogging service. We present multiple statistics that identify the evolution of blogging activity over the last few years in the portuguese blogosphere and compare it to existing studies on the worldwide blogosphere. Among other results, we conclude that the blogging activity has increased considerably over the years, both in the number of new blogs and in the number of posts created. We also observe growth in the usage of links in blogs and identify some opportunities for future research related to the evolution of the link ecosystem in the portuguese blogs.

Resumo

Ao longo dos anos, os blogues tornaram-se uma parte popular e influente da web que se distingue pelas suas principais características — são geralmente temáticos e organizados por comunidades, os seus conteúdos são apresentados por ordem cronológica e o seu interesse abrange tanto vários estudos sociais como diversos interesses económicos. Nesta dissertação, apresentamos uma caracterização da blogosfera portuguesa baseada num conjunto grande de blogues portugueses recolhidos a partir de várias fontes ao longo de um período de 5 anos.

Investigamos e aplicamos métodos que têm sido utilizados para a caracterizaçao da web e dos blogues. Utilizando amostras de blogues obtidas a partir de outras fontes, verificamos que uma quantidade satisfatória de blogues portugueses está coberta pela colecção. Analisamos as características desta recolha e estudamos a natureza dos blogues provenientes de diferentes fornecedores, de modo a identificar subconjuntos com diferentes características que possam trazer valor para efeitos de investigação. Estudamos o comportamento dos diferentes subconjuntos em várias características, incluindo a evolução do corpus destes ao longo do tempo e o número de entradas por blogue, e determinamos em que medida esta colecção pode ser representantiva da blogosfera portuguesa.

A nossa caracterização da blogosfera portuguesa é feita com base num subconjunto desta colecção que contém todos os blogues provenientes de um fornecedor português. A dimensão deste conjunto em particular permite-nos não só tirar algumas conclusões sobre a actividade dos blogues ao longo dos anos, mas também estudar o comportamento do conjunto de utilizadores de um determinado fornecedor. Desta forma, apresentamos diversas estatísticas que permitem observar a evolução da actividade nos blogues ao longo dos últimos anos na blogosfera portuguesa, comparando-a com os resultados de outros estudos existentes. Entre outros resultados, observamos que a actividade na blogosfera tem aumentado consideravelmente ao longo do tempo, tanto em número de novos blogues criados como novas entradas nos blogues. Também observamos a evolução da utilização de ligações em blogues e identificamos algumas oportunidades para estudos futuros relacionados com a evolução do ecossistema de ligaçẽs web entre os blogues nacionais.

Acknowledgements

For all the support and input given to this work since the beginning, I must thank both Cristina Ribeiro and Sérgio Nunes. I must also thank to the team behind SAPO for making the blog collection available for this reasearch.

Orlando Telmo Couto

"The Road goes ever on and on
Down from the door where it began.
Now far ahead the Road has gone,
And I must follow, if I can,
Pursuing it with eager feet,
Until it joins some larger way
Where many paths and errands meet.
And whither then? I cannot say."

J. R. R. Tolkien

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Blogs have become a popular and important form of communication over the web. Typically, they consist of special web pages with chronologically ordered entries, often with embedded links and occasionally a few images, usually being 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 [CK06]. In August 2008, the worldwide blogosphere was estimated to contain over 133 million blogs [Sif08]. Blogs have been the subject of many studies, both for social analysis and for diverse economical activities. Governments, corporations and traditional media seek to understand how to adapt to them and use them effectively, while citizens are using them as a tool to gain more voice in the world [JFJ+07].

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 portuguese people living in foreign countries. Although there is a significant number of portuguese bloggers that write in other languages, their blogs are hard to identify as being portuguese and are usually intended for the global blogosphere instead of the portuguese community exclusively. Although the first portuguese blogs have been around since the 90's, the portuguese community in general wasn't yet very familiar with the blogging activity by 2006 [Che08]. It was in this year that SAPO, a major portuguese ISP, launched a blogging service dedicated to the portuguese community.

1.1 Context

In the context of a research 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 solid and useful characterization of the portuguese blogosphere and how it has 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 particular service that is characteristic of a national community.

1.2 Motivation and Objectives

We survey and apply methods that have been used for web and blog characterization, while providing some insight on the results obtained by some authors of reference. Using samples of blogs obtained from different sources, we try to estimate the extent to which this collection can be representative of the portuguese blogosphere as a whole. The objective of this dissertation is to present multiple statistics that allow us to characterize the portuguese blogosphere and compare it to the results obtained from other studies.

1.3 Dissertation's Structure

Beside the Introduction, this dissertation contains 4 more chapters. In Chapter 2, we describe the state-of-the-art and present some studies of reference on characterizations of the blogosphere and local web communities. In Chapter 3, we characterize our blog collection and estimate the extent to which it can be representative of the portuguese blogosphere. In Chapter 4, we present a characterization of the portuguese blogosphere based on the results of our research. In Chapter 5, we summarize the main conclusions of this dissertation and identify some opportunities for future studies.

Chapter 2

Studies on Online Communities

The blogosphere has already been the subject of many studies over time. Several studies have provided characterizations of the blogosphere as a whole and also of smaller communities on the web.

2.1 Characterizations of the Blogosphere

Early studies on the blogosphere include the systematic description of the main characteristics of a blog [HSBW04] and a longitudinal content analysis of blogs published between 2003 and 2004 [HSKW06]. These authors discuss the differences between blogging and journalism and the influence of external events on the blogging activity. Three different sets of random english blogs were sampled and 22 features were analyzed, including some related to the blogger's characteristics like gender, age and occupation, and others related to the blog's contents. They concluded that most blogs were used as personal journals and that, despite the claims that links to other blogs are an essential characteristic of the blogosphere, they observed that links were less frequent than on the web and decreasing over time.

Also important for blogosphere characterization was the creation of the TREC Blogs06 Collection [MO06]. This blog test collection was created in order to provide a representative sample of the blogosphere for research purposes, with all the methodologies used in the process clearly stated. Among some statistics that resulted from the analysis of the collection, it could be observed that not all of the dates reported in XML feeds are to be trusted, since only 60% of the permalink documents had dates within the period of the crawling. The issue of splogs is also addressed and some patterns were detected that could allow for easier splog detection.

The link structure of the blogosphere is a characteristic that has been subject to many studies by itself. Early work on the spanish blogosphere [TRM03] was based on data from links collected from blogs on a daily basis. Analysis of the link popularity ranks concluded that the most popular links pointed to news websites and that only one tenth of the links pointed to other blogs. Tricas, Ruiz and Merelo also provided a model for measuring how often a blog could be found while traveling through the blogosphere from one link to another, allowing the identification of the most central blogs in the community.

Studying how the hyperlinks in blogs change over time and the most recent links for any given period, advancements were made on methodologies to identify emerging topics and interests in the blogosphere [CK06]. An "influential model" that measures a blog's importance based on a given topic was provided, as well as some experiments with the polarization of opinion to evaluate the impact of using link polarity in that model [JFJ+07]. Joshi et al. defined the concept of link polarity as the association between sentiments and links in blog posts, based on the text that surrounds the links. This study also showed the lack of appropriate data sets to fully test the new approaches presented.

Twitter[1], a popular microblogging tool which appeared in 2006, has also been subject to recent study. Although Twitter accounts are different in their nature, they also have some features that can be compared to blogs. One of the most relevant studies made on this service presents Twitter's growth rate, both in users and user posts, and provides insight into the user activity and geographical distribution [JSFT07]. This study aimed to detect topics of interest within user communities, based on a given user's friends and followers — which would correspond to the out-links and in-links between blogs that can be found in the blogrolls. It was possible to distinguish groups of users that often communicate with each other and to use term-based content analysis to identify main topics commented within those groups, allowing the detection of trends on user activity and motivations.

One of the major references for blogosphere characterizations are Technorati's[2] "State of the Blogosphere" reports. As of July 2006, 50 million blogs had been tracked [Sif06]. A thorough analysis of the corpus observed that the number of blogs in the blogosphere doubles every 6 months and that the portuguese language was used in 2% of the blogs. This study also relates the spikes of blogging activity to important events. In August 2008, 133 million blogs had been indexed by Technorati [Sif08]. Some interesting observations are the way brands and advertising entered the blogosphere and blogs are getting taken more seriously as sources of information.

[1] http://www.twitter.com
[2] http://technorati.com/

2.2 On Local Web Communities

Works made on local communities are very relevant for this study, since they are usually more focused on usage behaviors and how those communities have evolved. The portuguese web has already been subject to research [GS05], based on a set of documents that were hosted under the .pt domain or hosted in other domains but written in the portuguese language and with at least one incoming link from a page hosted under the .pt domain. Gomes and Silva concluded that there was a large number of websites with only one document and that 93% had less than 100 documents. Most of the documents had a URL length between 20 and 100 characters and 95.9% of them were text documents. An interesting observation was that most of the portuguese web pages didn't have any links to other portuguese sites.

OberCom's[3] flash report "Blogues e Blogosfera .pt" [Che08] presents a characterization of portuguese bloggers, based on a survey of a sample of the portuguese population in early 2006. 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. Among all blog users (readers and creators), 45% stated that their motivation was to seek for information on specific matters and another 27% sought for more information about recent news matters. When asked about the subjects on their most visited blogs, 41% had entertainment as their first choice and 22% others had journal-type blogs from a circle of friends. Blogs about politics had a great distance from those, in terms of popularity: only 4% of the bloggers had politics as a favorite subject. However, it becomes clear that the blogging activity in general hadn't been adopted by most of the portuguese internauts.

The persian blogosphere has also been subject of research [QRSA07] with a large dataset of persian blogs containing comments. Data was gathered by using a crawler and then an HTML parser to extract comment links and other information not available in some feeds. The primary analysis of their work was on comments, rather than the general characteristics of the blogs. Qazvinian et al. found that this collection contained an average of 3.6 comments per post and most of them originated from bloggers within the blogosphere. They also provide a model for comment distribution on a given post on the days after its submission. However, no study on the impact of the comments in the blogging activity was made in this work.

The blog collection used in our study has been previously used on a research about the application of the h-index in the evaluation of a blog's importance [Bra08] and on a research on the detection of blog trends [Pin08]. Branco and Pinto also collaborated to provide a brief characterization of this collection, containing 54,149 blogs with more

[3] http://www.obercom.pt/

Author	Dataset	Main characterization subjects
Tricas et al. [TRM03]	Spanish-speaking blogs (unspecified)	Link structure
Gomes, Silva [GS05]	46,457 portuguese websites	Number of documents, evolution and link structure
Herring et al. [HSBW04, HSKW06]	3 blog samples (203, 154, 100)	Blog content analysis, blogger characteristics, usage statistics, link and comment structure
Macdonald, Ounis [MO06]	TREC Blogs 06 (100,649 blogs)	Post distribution, link structure, term-based content analysis
Cohen and Krishnamurthy [CK06]	8,679 blogs	Blog link structure
Qazvinian et al. [QRSA07]	22,306 persian blogs	Comment distribution
Java et al. [JSFT07]	Twitter (1,348,543 posts)	Usage statistics and link structure
Branco [Bra08], Pinto [Pin08]	49,940 portuguese blogs	Blog usage characteristics, content analysis
Mishne, Glance [MG06]	36,044 blog comments	Comment usage and characteristics

Table 2.1: Comparison between some characterization works.

than 3 million entries, from which they filtered only the blogs within the time frame from January 2003 to December 2007. Their analysis was divided in two groups: social and evolutional. The social analysis focused on the distribution of the entries and blog usage over time. Evolutional analysis, on the other hand, focused more on the evolution of new blogs created and number of entries per day over time. They detected a sudden growth of activity during early 2006, followed by periods of faster and slower growth rates. Pinto considered content analysis as another group of characteristics, separating the studies on the blog link structure and post sizes from the evolutional analysis. They concluded that the average post size is around 160 words and that the portuguese bloggers link more often to themselves than to other blogs, with the number of links in posts increasing over time.

2.3 Conclusions

Table 2.1 gives an overview of the data sets used and the main focus of the mentioned studies. Most of them were based on blog collections with a corpus varying between 20,000 and 50,000 blogs. Macdonald and Ounis used a dataset of considerable dimensions with more than 100,000 blogs in the collection's corpus [MO06]. Considering this aspect, the collection used in our study stands above the average with more than 60,000

Characteristic	Authors
Title / url length	[GS05]
Blogging software used	[HSBW04]; [Pin08]
New blogs created through time	[Bra08]; [Pin08]; [JSFT07]; [Sif07]
Number of posts per blog	[Bra08]; [Pin08]
New posts created through time	[Bra08]; [JSFT07]
Hour of posting	[Bra08]; [MO06]; [Pin08]; [Sif07]
Number of posts per weekday	[Bra08]; [QRSA07]; [Pin08]
Word count per post	[HSKW06]; [Bra08]
Number of comments per post	[MG06]; [QRSA07]
Number of in/out links per blog	[Bra08]; [GS05]; [HSBW04]
Number of in/out links over time	[Bra08]; [Pin08]; [TRM03]
Number of in/out links within the community	[QRSA07]
Number of self-incoming links	[CK06]; [Pin08]
Posts with more incoming links	[Pin08]
Most referenced websites	[Pin08]; [Sif08]

Table 2.2: Some blog characteristics and authors of reference.

blogs in corpus. While some characterizations were focused in deep studies of very few characteristics, others provide a more generic analysis of multiple features of the blogging activity.

Some of the main characteristics that have been studied by researchers on this subject are presented in Table 2.2. There are two main groups of characteristics that can be identified — one related to the general characteristics of blogs and the other related to the analysis of links and link structures in the blogosphere.

Chapter 3

The SAPO Blogs Collection

In this chapter, we present the main characteristics of our dataset and try to estimate the extent to which this collection is representative of the portuguese blogosphere. Most authors apply their own criteria to the crawling process in order to create a representative dataset. However, this collection was built by a different party, which is why it is important to validate the dataset's contents. Our analysis is focused in two main features: the number of portuguese blogs covered by the collection and the distribution of blogs by service provider.

3.1 History of the 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]. This is a blogging service dedicated to the portuguese community that has gained popularity over time. The collection we use in our research was created in order to provide a good 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.

This dataset 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. The criteria used to insert a blog in the collection was simple: starting from selected portuguese blogs, the crawler would follow the links in those blogs in order to find other portuguese blogs with the assistance of a language analysis tool.

One of the most interesting features 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.

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

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. Besides, the crawler has been subject to many revisions over the years and it is, therefore, natural that some inconsistencies may be found in the collection.

3.2 Blog Coverage

In order to provide a realistic characterization of the portuguese blogosphere, estimating if the collection is representative of this community is a matter of the utmost importance. 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.

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.

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. Then, we defined Blog Ratio as the percentage of blogs matching our collection within all blogs found with Blog Search.

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 non-portuguese blogs written in Portuguese, such as brazilian blogs, for example.

3.2.1 Calculating Blog Ratios

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 existed in the collection. We defined the Blog Ratio as the percentage of blogs matching our collection within all blogs found with Blog Search.

The biggest challenge while retrieving the Blog Ratios within our collection was to decide which queries to use, since we needed to define queries that would be inherent

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

Query	Posts	Blogs	Matches	Blog Ratio
25 liberdade revolução cravos	100	80	25	31.3%
arguido bragaparques	100	60	25	41.7%
bento manuel galrinho	100	86	50	58.1%
excomunhão tarcísio cónego	68	64	39	61.0%
fiama hasse brandão	89	69	37	53.6%
natal feliz festas santo deseja próspero	100	94	44	46.8%
prado coelho eduardo epc	100	78	44	56.4%
psd menezes directas eleições	100	58	21	36.2%
referendo aborto abstenção despenalização	100	84	47	56.0%
santana lopes chelsea mourinho interrompido	35	32	13	40.6%
salazar cunhal	100	88	37	42.0%
Total	992	793	382	48.2%
Average	90.2	72.1	34.7	47.6%

Table 3.1: Blog Ratios for popular topics.

to the portuguese culture, like important events or words that are less common in the Brazilian Portuguese language. Pinto's work on the detection of blog trends and popular topics [Pin08] was an important step in the study of blog contents within the portuguese blogosphere so, in our first approach, we decided to query for some of the most popular topics detected during 2007, focusing on queries that contained only words in Portuguese.

The queries we used and the respective Blog Ratios from our first analysis are shown in Table 3.1. For each query, we present the number of posts retrieved from Google Blog Search, followed by the number of unique blogs detected among those posts and, in column "Matches", the number of these unique blogs that exist in the collection. The Blog Ratio is the percentage of blogs that exist in the collection, in relation to the number of unique blogs found. Finally, we present the total and average numbers of blogs found for all queries and the corresponding Blog Ratios.

The results were satisfactory: an average of 47.6% of the blogs retrieved from Blog Search queries belong to our collection. However, some factors must be observed while analyzing these results. Google Blog Search often returns posts from foreign blogs (from portuguese-speaking countries) and posts from news sites, which are difficult to detect and remove in automatic routines. Another aspect that must be noticed is that Google's API limits the results to 100 posts, retrieving the most popular according to their ranking system. This means that most of Table 3.1's results are biased on the most popular posts found. Since we are estimating the coverage of the entire portuguese blogosphere by the SAPO Blogs collection, it is important to eliminate the bias on popularity by using queries that produce less than 100 results in Blog Search.

Query	Posts	Blogs	Matches	Blog Ratio
25 liberdade revolução cravos grândola	98	97	36	37.1%
arguido bragaparques caso carmona lisboa depor	13	12	7	58.3%
bento manuel galrinho benfica adeus	45	34	14	41.2%
excomunhão tarcísio cónego	68	64	39	61.0%
fiama hasse brandão	89	69	37	53.6%
natal feliz festas santo deseja próspero ano novo votos amigos familiares	97	85	27	31.8%
prado coelho eduardo epc crítico literário	42	32	19	59.4%
psd menezes directas eleições rumo mudou	34	20	12	60.0%
referendo aborto abstenção despenalização vinculativo eleitores	74	64	40	62.5%
santana lopes chelsea mourinho interrompido	35	32	13	40.6%
salazar cunhal concurso rtp grandes portugueses estado novo pessoa	91	61	38	62.3%
Total	686	570	282	49.5%
Average	62.4	51.8	25.6	51.6%

Table 3.2: Blog Ratios for popular topics with refined queries.

3.2.2 Refining the Queries

In order to address the problem associated with Google's ranks, we decided to repeat our calculations for blog coverage with longer, more refined queries. Our refinements consisted in adding more portuguese words related to those already in the query, until the number of posts retrieved for each query was below the 100 post limit. The new queries are presented in Table 3.2, as well as their respective results. An average of 51.6% of the blogs retrieved with these queries could be found within our collection. In some queries, we observed a big improvement on the number of matches with blogs from our collection, which confirms that some blogs were previously ignored for being "less popular". Therefore, we can state that refining our queries allows us to obtain more realistic results for the blog coverage ratios.

3.2.3 Validating Results

We have studied the collection's coverage of the portuguese blogs based on a set of queries that were known to have been popular topics, since they were expected to provide positive results. This test allowed us to verify that, generally, half of the portuguese blogs retrieved that mentioned those topics are contained in this collection. Although these results were considered satisfactory, we needed to verify them by testing the collection with different topics.

Query	Posts	Blogs	Matches	Blog Ratio
abrupto pacheco pereira psd opinião política contra aborto referendo leitores	78	15	10	66.7%
bloco esquerda liberalização referendo aborto opção mulher	56	48	30	62.5%
boa vida férias descansado trabalho praia sol gaja	67	51	28	54.9%
casa da música concerto clubbing optimus	88	36	14	38.9%
casamento homossexual psd ferreira leite família procriação	99	79	30	38.0%
cinha jardim castelo branco júlia pinheiro	75	22	11	50.0%
facto governo socialista sócrates magalhães escola criança educação ministra vergonha demissão política	90	19	9	47.4%
fantasporto festival vampiros	44	37	15	40.5%
gajas boas noite sair discoteca engate	73	64	33	51.6%
luciana abreu programa sic manhã roupa novo look	78	39	16	41.0%
noite lisboa bairro alto tomar copo chiado	57	37	16	43.2%
triste vidinha desemprego combustíveis	77	35	16	45.7%
Total	882	482	228	47.3%
Average	73.5	40.2	19	48.4%

Table 3.3: Blog Ratios for other topics.

Provider	Blogs	%
SAPO	120,178	83.0%
Blogger	23,588	16.3%
Weblog	718	0.5%
Other	307	0.2%
Total	144,791	100%

Table 3.4: Blogs in collection, by service provider.

We used a new set of queries related to more recent events and some informal linguistic expressions. In order to avoid a potential bias on given topics, we tried to ensure that the new queries covered different aspects of the portuguese reality — controversial political issues, social events, lifestyle, etc. These queries and respective results can be observed in the Table 3.3. We obtained an average of 48.4% blog coverage, below the coverage obtained for the previous set but not for a significant difference. Looking globally at the results from both query sets, we can state that half of the portuguese blogs found within Google's Blog Search engine can also be found within our collection.

3.3 Blog Service Providers

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

We retrieved the number of blogs from each server to verify this situation, which can be seen in Table 3.4. From all blogs in the collection, 83% are hosted by the SAPO Blogs, while 16.3% are hosted by Blogger[3]. Other blog service providers have a marginal representation, where Weblog[4] stands out the most.

We used the Google Blog Search results from our previous study to measure the number of blogs from each provider among them, in order to compare with our collection. Although this was a very small sample of 1587 blogs, the results presented in Table 3.5 show a different reality than the one presented in our collection. Most of the retrieved blogs are hosted by Blogger, which is Google's own blogging service and known to be extremely popular among the worldwide blogosphere [HSBW04]. The SAPO Blogs service is the second most popular provider and has a good expression, but far from our collection's numbers. These results also indicate that 15.9% of the portuguese bloggers

[3] http://www.blogspot.com
[4] http://weblog.com.pt/

Provider	Blogs	%
SAPO	204	12.9%
Blogger	1,131	71.2%
Weblog	30	1.9%
Wordpress	76	4.8%
Other	146	9.2%
Total	1,587	100%

Table 3.5: Blogs from Blog Search results, by provider.

use other services, with Wordpress[5] standing out, while our collection barely acknowledges them.

The difference between the results from Tables 3.4 and 3.5 rises a question about the results from our blog coverage analysis: if there is such a great 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 apparent contradiction, we decided to calculate the number of posts per blog for each service provider. The results are presented in Table 3.6. Work on a previous version of this collection observed that blogs hosted by Blogger had more posts than the ones hosted by SAPO [Pin08]. However, the difference between the average number of posts from blogs hosted by SAPO and the other providers is larger than expected. Besides, we found many blogs without posts while retrieving the number of posts for all blogs.

3.3.1 Empty Blogs

We decided to retrieve all the blogs without posts in our collection and try to evaluate the impact of those empty blogs in our collection. Table 3.7 presents the number of empty blogs that originate from each provider. The total number of empty blogs is 82,961, which represents 57% of our collection. The number of empty blogs hosted by SAPO stands at nearly 67% of all blogs from this provider, while there are fewer empty blogs from other

[5] http://wordpress.com

Provider	Blogs	Posts	Average
SAPO	120,178	818,797	6.81
Blogger	23,588	2,630,178	111.50
Weblog	718	59,890	83.41
Other	307	53,761	175.12
Total	144,791	3,562,626	24.61

Table 3.6: Number of posts in all blogs by provider.

Provider	Total Blogs	Empty Blogs	%
SAPO	120,178	80,133	66.68%
Blogger	23,588	2,645	11.21%
Weblog	718	132	18.38%
Other	307	51	16.61%
Total	144,791	82,961	57.30%

Table 3.7: Number of empty blogs in collection by provider.

providers. Looking at this data from another perspective, we can see in Table 3.8 that 96.6% of the empty blogs in this collection are hosted by SAPO.

We have gathered a sample of the empty blogs and inspected them manually. Many of the blog URLs we tested — except from those hosted by SAPO — were blogs that had been deleted or couldn't be found for a variety of reasons. We also found some blogs that had posts, but they may have not be found due to some problem during the collection's crawl. However, the empty blogs from these servers represent only 3.4% of all the empty blogs and have no significative impact in our collection's representativity.

On the other hand, we found that most of the empty blogs hosted by SAPO actually existed as empty blogs. Unlike other blog service 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.

The empty blogs in our collection provide very little information. Apart from the blog's name and URL address, there is no data that could allow us to track when those blogs were created. Since these blogs do not contain most of the common features from a typical blog, we decided to filter them out of our research.

3.3.2 Posts in Blogs

Without the empty blogs, we can now estimate the proper distribution of blogs and posts in our collection by each service provider. As can be seen in Table 3.9, nearly 65% of the

Provider	Empty Blogs	%
SAPO	80,133	96.60%
Blogger	2,645	3.19%
Weblog	132	0.16%
Other	51	0.05%
Total	82,961	100 %

Table 3.8: Distribution of empty blogs by provider.

Provider	Blogs with posts	%
SAPO	40,045	64.77
Blogger	20,943	33.88
Weblog	586	0.95
Other	256	0.40
Total	61,830	100

Table 3.9: Number of blogs with posts by service provider.

blogs in the collection are hosted by SAPO. These numbers, however, remain inconsistent with our previous Blog Search results from Table 3.5. We also decided to investigate the number of posts per blog without the empty blogs affecting our calculations.

Table 3.10 presents the absolute and average numbers of posts per blog from each provider in the collection. The collection contains an average 57.6 posts per blog, but there is a large discrepancy between blogs hosted by SAPO and the ones hosted by other providers. Although the number of blogs hosted by Blogger is nearly half the number of those hosted by SAPO, they contain 3 times more posts than the latter. Since we already filtered the empty blogs, we needed to find an explanation for such a large difference between numbers of posts.

If we take into account the massive amount of blogs that are registered but are never used within the SAPO Blogs service, we can assume that there is also a large amount of blogs that contain very few posts. We retrieved the number of blogs from each provider that contained only one post, which are presented in Table 3.11, and we found that 21.6% of the blogs in the collection had only one post. However, the most important observation is the difference between the blogs hosted by SAPO and the other blogs in the number of blogs with one post. This is directly related to the difference between the average numbers of posts per blog we previously found. However, such discrepancies must be related with the collection's nature. Since this collection is based on SAPO's entire dataset and the blogs from other providers were the result of a crawling process, blogs with many posts would be more likely to be found and included in the collection than blogs with fewer posts.

Provider	Blogs	Posts	Average
SAPO	40,045	818,797	20.45
Blogger	20,943	2,630,178	125.59
Weblog	586	59,890	102.20
Other	256	53,761	210.00
Total	61,830	3,562,626	57.62

Table 3.10: Average number of posts per blog by provider.

Provider	Blogs	Only 1 post	%
SAPO	40,045	12,511	31.24%
Blogger	20,943	744	3.56%
Weblog	586	51	8.70%
Other	256	42	16.41%
Total	61,830	13,348	21.59%

Table 3.11: Blogs with only 1 post by provider.

We analyzed the number of blogs that have reached a specified number of posts. Figure 3.1 depicts the different behavior of blogs hosted by SAPO and Blogger. The curves for both providers are very different: it is clear that half of the blogs hosted by SAPO have ended at 4 posts or less, but only a few blogs from Blogger in our collection have ended below that limit. This confirms that the crawler was able to retrieve only a portion of the blogs with many posts from the other providers, explaining the difference between the average number of posts per blog in different providers.

It is also important to notice that, as the number of posts increases, as shown in Figure 3.1, the curves from both providers have increasingly similar behaviours - which might indicate that, if we had the entire collection from other providers, the curves would be similar to SAPO's from the beginning. Blogger's curve also drops suddenly at the number of 25 posts, breaking an apparently linear trend. 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.

3.3.3 Other Characteristics

We found that the blogs in the collection that were hosted by Blogger correspond to a specific set of blogs from that provider that contain a higher average number of posts than the blogs hosted by SAPO. For this reason, we decided to analyze other characteristics from the blogs hosted by these providers, in order to compare the behaviors in the different sets of blogs. We left the blogs from other providers out of this analysis, due to the marginal representation they have in the corpus of the collection.

We retrieved the date of the first post from each blog, in order to identify when the blogs from SAPO and Blogger were created. Figure 3.2 presents the number of blogs created per month since January 2005. It is clear that the blogs from these two providers were created in different time periods, indicating that the crawler found many blogs from Blogger that originated between late 2005 and early 2007, while most of the blogs from SAPO were created after July 2007.

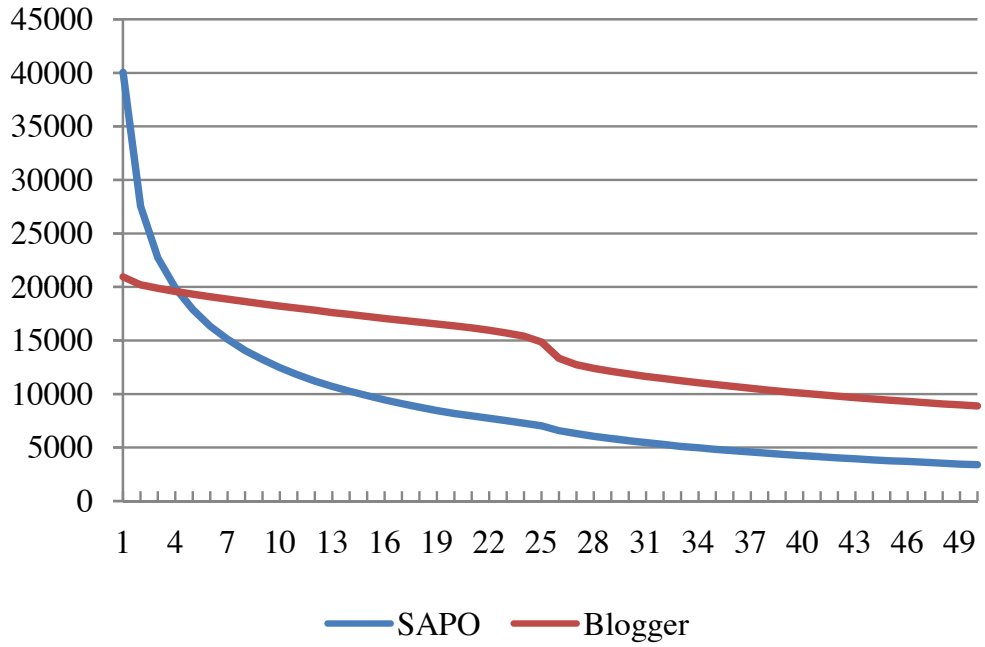

Figure 3.1: Number of blogs with *n* or more posts, by provider.

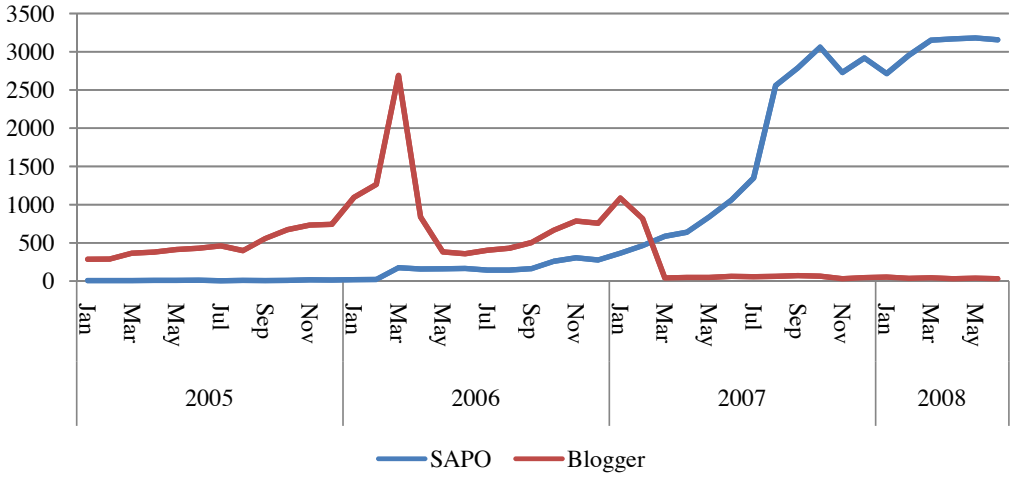

Figure 3.2: Number of blogs created per month, by provider.

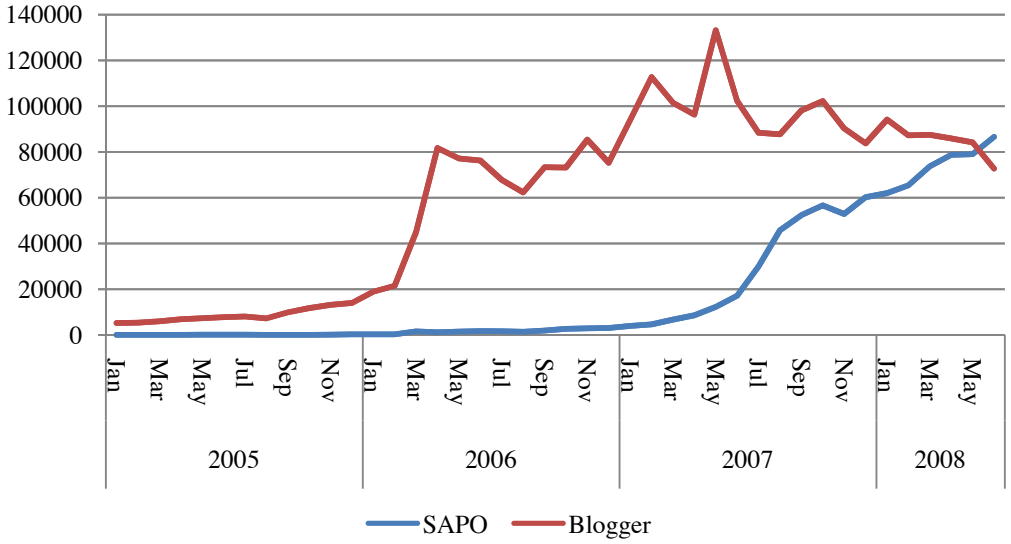

Figure 3.3: Number of posts created per month, by provider.

The SAPO Blogs service was launched in March 2006 and had a slow adoption rate during the first year of activity. We cannot draw any conclusions about the growth of the corpus in blogs hosted by Blogger, since all the information we have is the date when this set of blogs were created. Therefore, using these blogs to represent the growth of the portuguese blogosphere would lead to erroneous results.

On the other hand, it is interesting to analyze the behavior of this set of blogs in terms of posting activity and compare them to the blogs hosted by SAPO. We retrieved the dates from all posts in blogs from the collection and counted the number of posts created per month from each provider, as shown in Figure 3.3. Regardless of the period when most of the blogs hosted by Blogger were created, their posting activity remains very high throughout the timeline. This reinforces the statement that this set contains blogs not only with high post counts, but also featuring a vast life span. Another observation is that the posts from blogs hosted by SAPO appear to grow at a faster rate than the rate of creation of new blogs, indicating that the bloggers have increased their posting activity through time.

Another characteristic we observed by analyzing the post dates was the time of day when posts were created. Figure 3.4 depicts the distribution of posts per hour from the blogs hosted by SAPO, in comparison to the blogs hosted by Blogger. In general, the posting activity is very similar between the two groups, although blogs from the Blogger service have their posts distributed more evenly over the time of day. This shows us that the posting habits don't differ much between users of different blogging services.

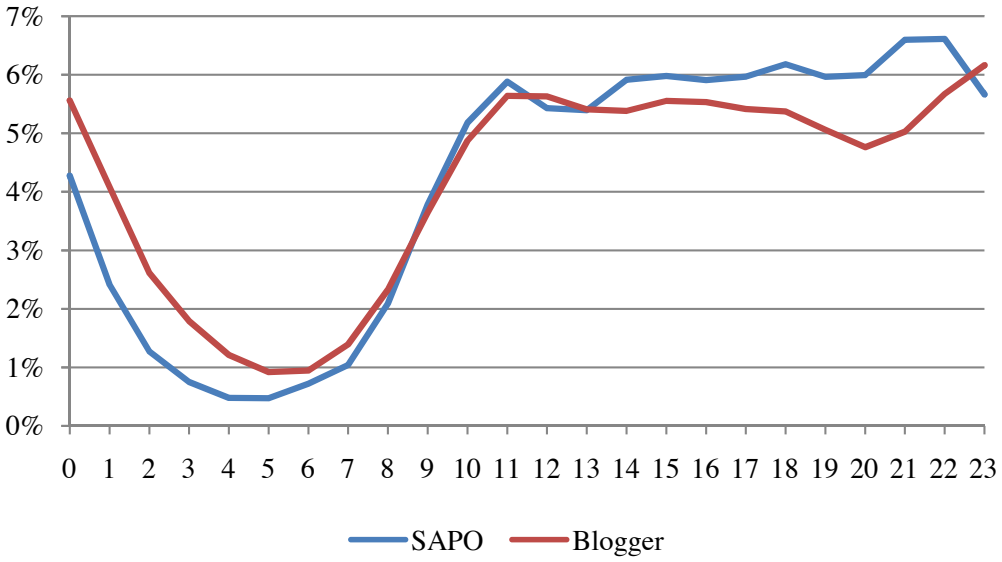

Figure 3.4: Distribution of posts created per hour, by provider.

3.4 Conclusions

After studying our collection's representativeness, we concluded that it presents a good coverage of the portuguese blogosphere, since half of the blogs provided from Google's Blog Search results can be found within the collection. Although the sources for our blogs aren't representative of the reality in terms of blog service usage, specially considering that Blogger is a highly popular provider with small coverage in this collection, we also observed that the blogs hosted by SAPO sport a behavior that is similar to the rest of blogs in the collection.

We concluded that the blogs hosted by Blogger in the collection may not be used as a reference for many features in a characterization of the portuguese blogosphere. However, they compose a valuable set of blogs with high post counts and a considerable life span that can be used for many research purposes such as the trend detection or the evolution of the link ecosystem in the blogosphere.

On the other hand, the blogs hosted by SAPO form a complete dataset from a portuguese blog service provider that, as we observed, sport a similar behavior to other known portuguese blogs. This indicates that this subset in our collection is representative of the portuguese blogosphere in a smaller scale.

Chapter 4

Characteristics of the Portuguese Blogosphere

As observed in the previous chapter, the blogs hosted by SAPO form a subset of the SAPO Blogs collection that is representative of the behavior in the portuguese blogs in general. On the other hand, blogs retrieved from other providers have very specific characteristics that could bias our results in the characterization of the portuguese blogosphere. Taking that into account, this characterization will be based exclusively in the set of blogs from the collection hosted by the SAPO Blogs service. Although this is a relatively recent provider when compared to others, we benefit from having its entire dataset, which allows us to draw conclusions on the growth of the portuguese blogosphere over the last few years.

4.1 Evolution of Blog Growth

The growth of the blogosphere is often evaluated by the number of blogs created over time. Figure 4.1 depicts the number of new blogs created per month since the launch of the SAPO Blogs service. As we can see, during the first year of activity, the number of new blogs created had a very slow growth. It was during the year of 2007, specially during the second half, that the activity in this service really took off, with more than 2,500 blogs being created per month. However, the creation of new blogs per month seems to stabilize over time.

Another way to look at the same data is by considering the total number of blogs created over time, as can be seen in Figure 4.2. The difference between growth during 2006 and growth after July 2007 might be due to a better awareness of the blogging activity by the portuguese people, which was very low when the SAPO Blogs service was

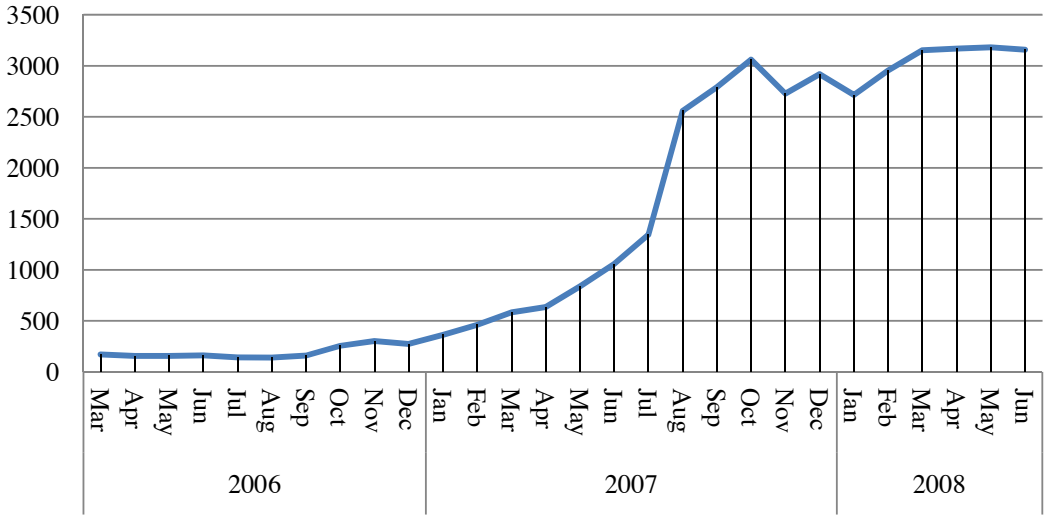

Figure 4.1: Number of blogs created per month.

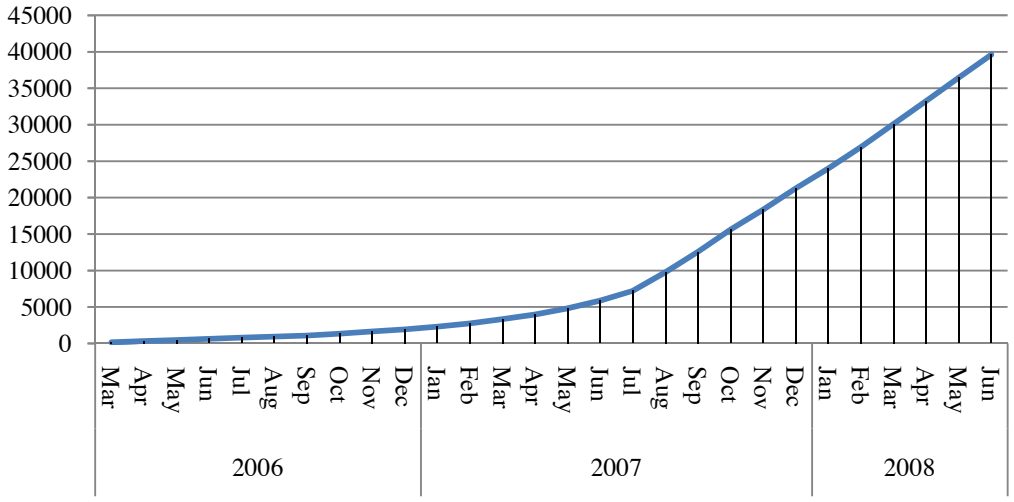

Figure 4.2: Total number of blogs through time.

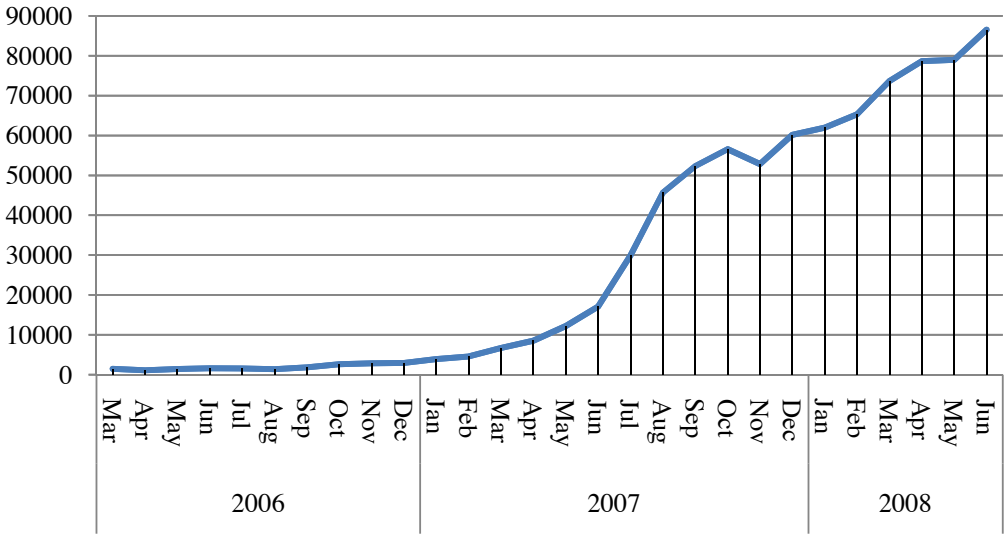

Figure 4.3: Number of new posts created through time.

launched [Che08]. Part of this growth might also be a result of marketing campaigns that raised the awareness and popularity of this service. In the long term, the trend observed in this set of blogs follows the same growth pattern observed in the worldwide blogosphere [Sif07].

4.2 Blog Post Activity

The posting habits of the portuguese community and how they have evolved are also interesting to analyze. We retrieved all post dates in order to determine the number of posts created per month and analyze the evolution of posting activity over the timeline, as shown in Figure 4.3. The observed behavior is similar to the evolution of the creation of new blogs per month during the first year of activity. However, during the second half of 2007, the posting activity started increasing at a much higher rate. This indicates that portuguese bloggers started posting more frequently over time, which is an interesting characteristic to analyze.

4.2.1 Number of Posts per Blog

With both the number of blogs and posts increasing over time, we decided to evaluate the number of posts per blog and how that number has evolved. Figure 4.4 presents the number of blogs per total number of posts submitted. It can be seen that most blogs are one-time experiences, since most of them contain less than 5 posts. However, there is also

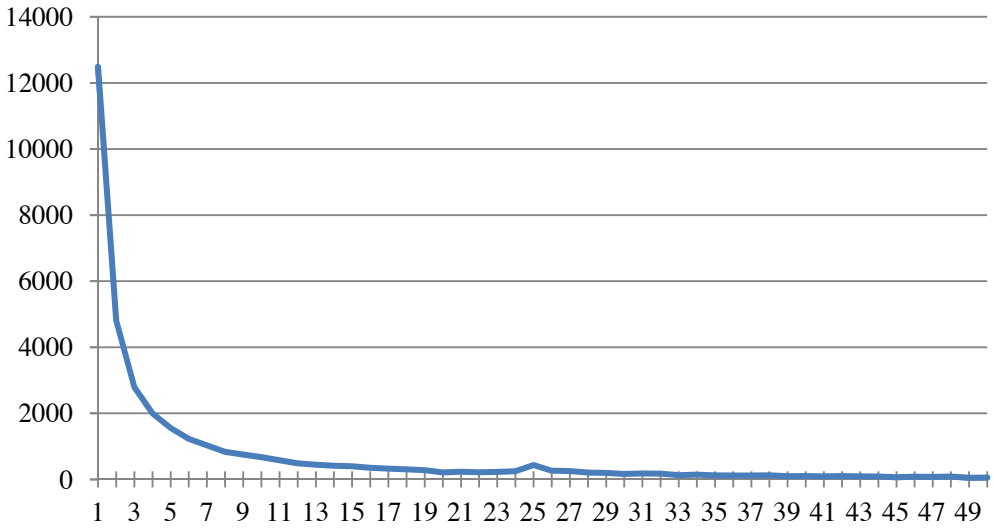

Figure 4.4: Number of blogs per number of posts.

a long tail of blogs with a high number of posts. This accounts for an average 20 posts per blog in the entire set, as previously seen in Table 3.10.

We also detected a minor inconsistency in the number of blogs with 25 posts. We suspect that this anomaly might be related with the construction of the dataset, and not a characteristic of the blogosphere — similarly to what we found in Section 3.3.

We retrieved the cumulative numbers of blogs and posts created over time, in order to estimate the growth of the posting activity. Figure 4.5 depicts the evolution of the average number of posts per blog over time, obtained by calculating the absolute numbers of blogs and posts created since the beginning up to each month in the timeline. 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.

4.2.2 Hour of Posting

As we previously observed in Figure 3.4, the time when posts in portuguese blogs are created doesn't differ much between different providers. However, we wanted to analyze how those habits have changed over the course of the years. Figure 4.6 shows a comparison of the post distribution according to the reported hour of posting during the different years covered by the collection. The general behavior has remained very similar over the

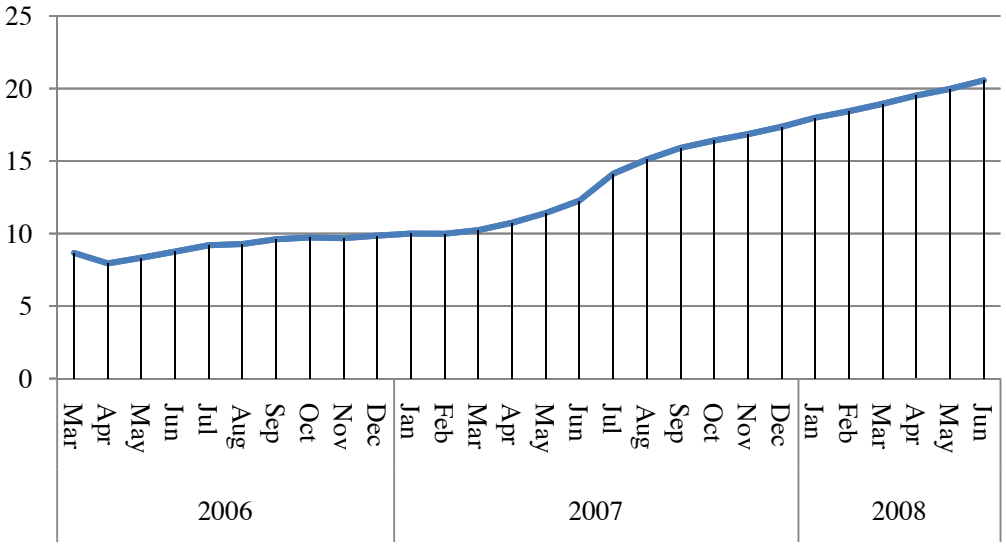

Figure 4.5: Average number of posts per blog through time.

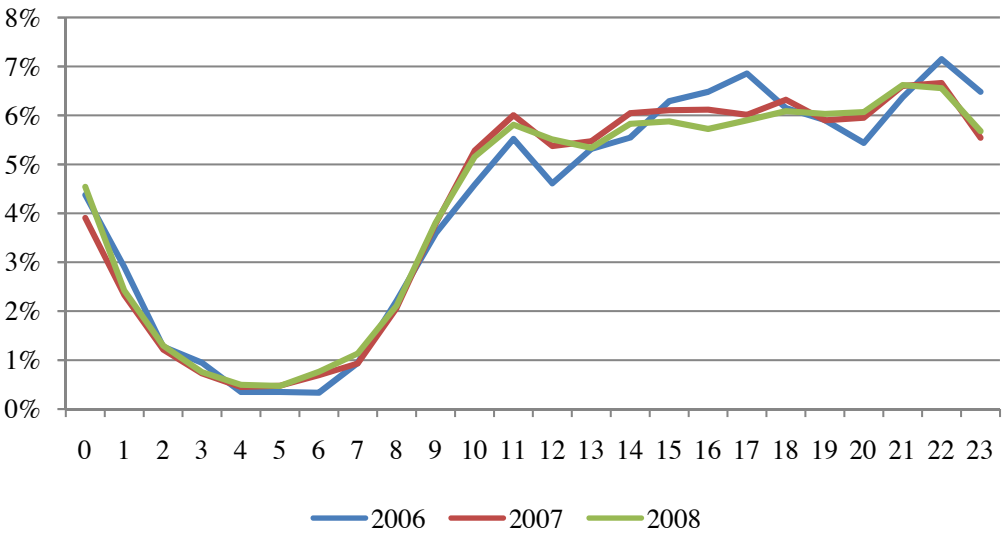

Figure 4.6: Post distribution per hour over the years.

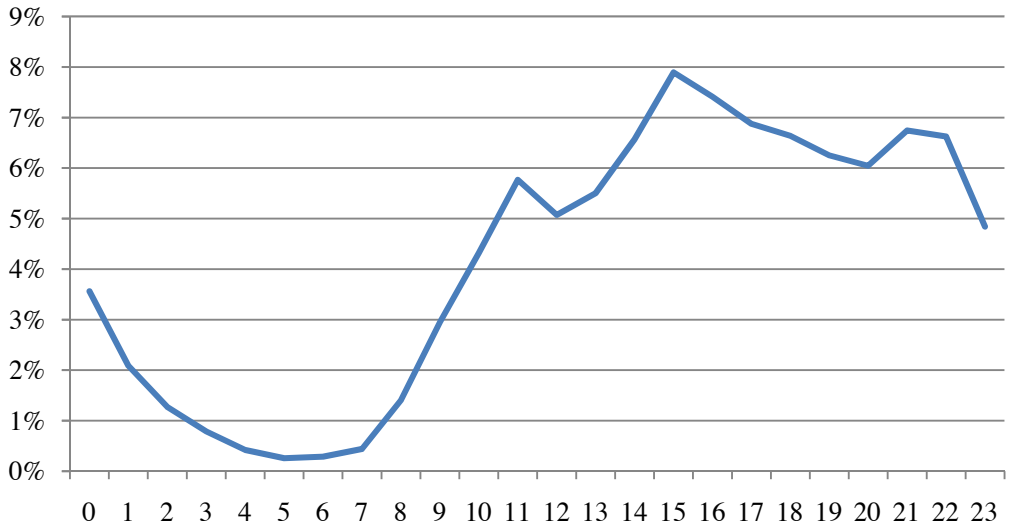

Figure 4.7: Distribution of new blogs created per hour.

years, with an apparent trend for the posts to be submitted more evenly over the time of day as the blogosphere grows and the number of posts per blog increases. This is reinforced by the difference of distribution during the year of 2006, when there were very few blogs, in comparison to the later years.

Looking globally at these results, we can observe that 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, and there is also a clear peak of activity after dinner.

Regarding the results obtained back in Figure 3.4, we can consider the set of blogs hosted by Blogger — characterized by their high life span and post count — as a possible trend to be followed by the portuguese blogosphere as blogs start to mature. This trend is consistent with the results obtained from other studies made on the worldwide blogosphere, with similar periods of higher and lower activity [MO06].

4.2.3 Hour of First Post

We analyzed the distribution of posts by the time of day and how it evolved over the years. An interesting characteristic to analyze and compare with these results is the distribution of the number of blogs created according to the time of day.

For that matter, we retrieved the distribution of new blogs created according to the time registered in the first post of each blog, as depicted in Figure 4.7. The results show that new bloggers are more compelled to write their first post after lunch time or later at

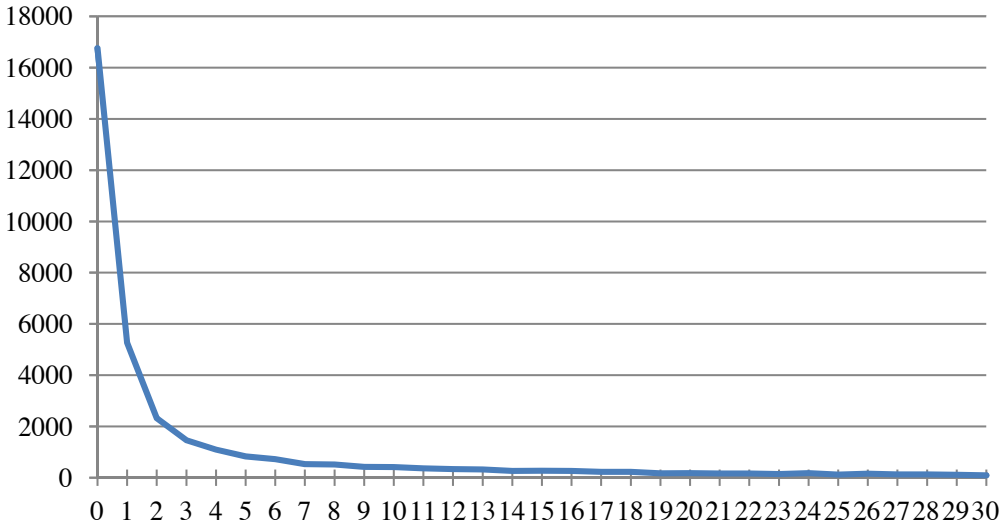

Figure 4.8: Number of blogs per total number of links used in posts.

night, indicating a tendency to wait for a period of free time to write their first post. This reinforces the statement that, as their activity increases, bloggers tend to write more in different times of day.

4.3 Link Usage in Blogs

The usage of links in blogs is a characteristic of considerable importance to the understanding of the blogosphere. We retrieved the total number of links used in all posts from each blog and observed that most blogs use only a few links during their life span, as can be seen in Figure 4.8. However, there is also a long tail of blogs with high numbers of links.

In order to analyze the evolution of link usage in blogs, we retrieved the number of links created over the years, as depicted in Figure 4.9. During the first year of activity, very few links were used. The growth in link usage appears to follow the same trends seen in the number of new blogs and posts created — similar growth of activity during the second half of 2007 and similar peaks of activity can be observed. However, link activity during the months of March and June 2008 appears to be inconsistent with the evolution observed during the rest of the timeline. We were unable to address the reasons behind this apparent anomaly but, considering that the estimated number of links during May 2008 seems consistent with the rest of the timeline, we believe that these two peaks of activity may be due to a problem that occurred during the collection's retrieval[1].

[1] We contacted the creators of the dataset, but they also had no explanation for this situation.

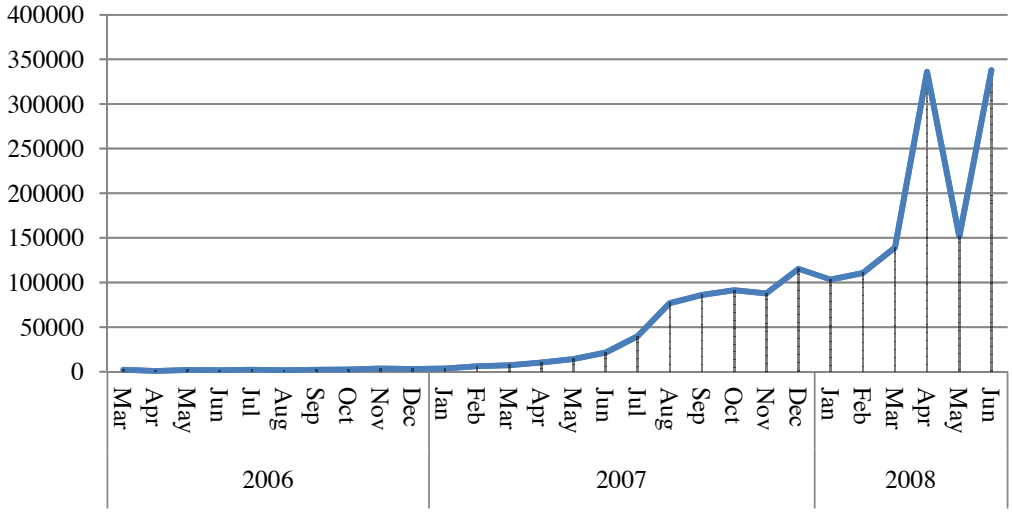

Figure 4.9: Number of links created through time.

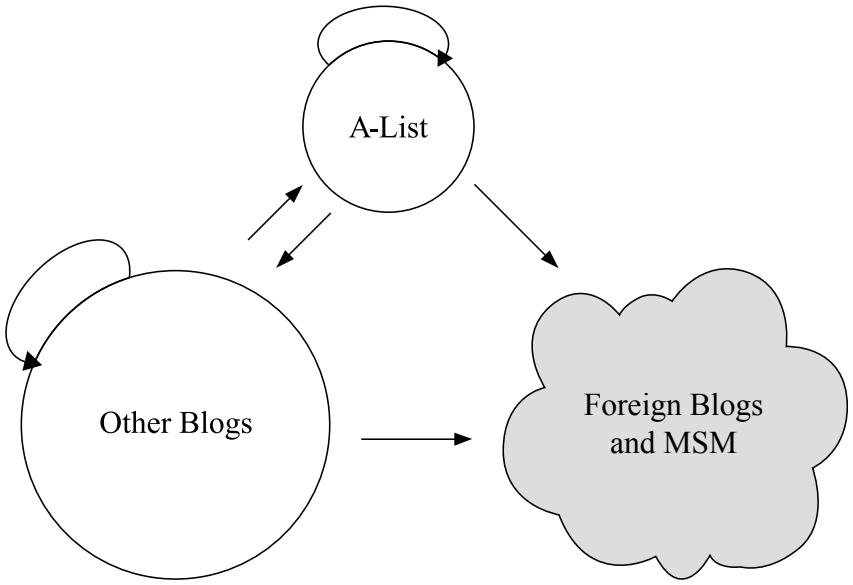

Figure 4.10: Possible model for the link ecosystem in the portuguese blogosphere.

We observed that the usage of links in the portuguese blogosphere has increased over time, appearing to follow the trends of growth in posting activity. Future research should analyze the relation between these two activities in order to determine the number of links per posts and how it evolved over the years.

The analysis of links in the blogosphere opens a window of opportunities for future research, such as the detection of the most influential blogs in the community and the structure of the link ecosystem in the portuguese blogosphere. Figure 4.10 presents a possible model for the portuguese link structure that would allow to understand the behavior in different groups of the blogosphere and detect the most influential sources for each group. Each arrow represents the number of links from each set of blogs that point to the most influential blogs (often called "A-List") and to other blogs from the portuguese blogosphere, as well as blogs outside the portuguse blogosphere and other sites from the mainstream media (MSM).

4.4 Conclusions

We observed that the portuguese blogosphere has increased over the years, in numbers of new blogs and posts created. In June 2008, portuguese blogs had an average number of 20 posts, nearly doubling in relation to same period in the previous year. Most blogs are created after lunch or after dinner but, as the blogging activity increases, portuguese bloggers tend to post in more evenly distributed periods of the day. The link usage in blogs, although infrequent during the first years of activity, has increased as the portuguese blogosphere started to mature.

Chapter 5

Conclusions

We observed that most studies of reference on blogosphere characterizations are divided between works that provide superficial analysis of multiple characteristics and works that delve deep into specific features of the blogging activity. These works provided valuable insight on the possibilities for our research. However, we had to draw a line to decide when to stop detailing the analysis of a specific characteristic and begin researching another.

A considerable amount of time was spent during the analysis of the collection's representativeness and part of this could have been overlooked, leaving more time for the study of other characteristics. However, we believe that the extent of research spent in understanding the collection has payed off in the end. We were able to detect groups of blogs with different characteristics and identify multiple opportunities for research that could take advantage of this collection. More importantly, the knowledge obtained with this analysis allowed us to filter blogs from the collection that would otherwise bias the results in our characterization.

We were able to provide multiple statistics that characterize the blogging habits among portuguese bloggers and identify opportunities for future research. Given the volumes of data to be processed, retrieving and processing the information was a slow process that required both time and patience — sometimes, a few days were spent to obtain data for a single characteristic.

Our characterization provides not only new information about the portuguese blogosphere, it also presents statistics regarding the overall behavior among users of a particular service. We compare different sets of blogs and analyze their evolution over the last few years. Our characteristics include numbers of blogs and posts per month, post rate over time and the evolution of link usage. We also identify and present opportunities for future work on the portuguese blogosphere.

A paper presenting some of the results obtained with our research was accepted for publication in the 3rd Int'l AAAI Conference on Weblogs and Social Media[1] [CRN09]. The paper, entitled "Characterizing the Portuguese Blogosphere", is centered in the representativeness of the collection and includes some preliminary results from the characterization study.

5.1 Summary of Results

We examined our collection's representativeness and observed that it presents a good coverage of the portuguese blogosphere, based on results obtained with Google Blog Search. The collection is formed by two main sets of blogs hosted by two different providers — Blogger and SAPO. Although we found that this isn't representative of the blog service usage in the portuguese blogosphere, we found similar blogging habits between different service providers.

We observed that the blogs hosted by Blogger in the collection compose a valuable set of blogs with high post counts and a considerable life span that can be used for many research purposes such as the trend detection or the evolution of the link ecosystem in the blogosphere. However, they have many distinct features due to the method used during the creation of the dataset that impede us from using them in a reliable characterization of the portuguese blogosphere. The main reasons behind this are that these blogs originate from a very specific period in time and that this particular set lacks most of the smaller blogs from this provider.

On the other hand, the blogs hosted by SAPO form a complete dataset from a portuguese blog service provider that, as we observed, sport a similar behavior to other known portuguese blogs, specially concerning the posting ratios over time. Therefore, we concluded that this subset from our collection can be considered representative of the portuguese blogosphere in a smaller scale. Besides, it allows us to identify interesting characteristics regarding overall blogging behavior among all bloggers using a particular service that is representative of a national community.

Most of the blogs were found to be one-time experiences or containing less than 5 posts. However, we observed that the portuguese blogosphere had considerable growth over the years, not only in the creation of new blogs but also in their posting activity. The average number of posts per blog nearly doubled in the last year of the timeline and the number of links used in blog posts followed the same trend of the general blogging activity.

[1] http://www.icwsm.org/2009/cfp.shtml

5.2 Future Work

A characterization of the blogosphere is never complete, as new opportunities for research appear for each result that is obtained. Analyzing the rate at which new posts are created in blogs could give us insight and allow us to observe the behaviors during different stages of a blog's life cycle. Another interesting characteristic that should be subject of further analysis is the link usage in portuguese blogs. This analysis could lead to the detection of the most influential bloggers and observation of the behaviors from different groups within the community. Identifying the relationships between blogs within the portuguese blogosphere and how they relate with foreign blogs and other types of media might provide insight to the detection of authorities among portuguese bloggers.

References

[Bra08] José Mário Branco. Aplicação do H-Index em Blogues. Master's thesis, Faculdade de Engenharia da Universidade do Porto, July 2008.

[Che08] Rita Cheta. Bloguers e Blogosfera .pt. Technical report, OberCom, 2008.

[CK06] E. Cohen and B. Krishnamurthy. A Short Walk in the Blogistan. *Computer Networks*, 50(5):615–630, 2006.

[CRN09] T. Couto, C. Ribeiro, and S. Nunes. Characterizing the Portuguese Blogosphere. In *Proceedings of the 3rd Intl AAAI Conference on Weblogs and Social Media*, January 2009. Accepted for publication.

[GS05] D. Gomes and M.J. Silva. Characterizing a National Community Web. *ACM Transactions on Internet Technology (TOIT)*, 5(3):508–531, 2005.

[HSBW04] S.C. Herring, L.A. Scheidt, S. Bonus, and E. Wright. Bridging the Gap: a Genre Analysis of Weblogs. In *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*, pages 101–111, 2004.

[HSKW06] S.C. Herring, L.A. Scheidt, I. Kouper, and E. Wright. A Longitudinal Content Analysis of Weblogs: 2003-2004. *Blogging, Citizenship, and the Future of Media*, pages 3–20, 2006.

[JFJ+07] A. Joshi, T. Finin, A. Java, A. Kale, and P. Kolari. Web (2.0) Mining: Analyzing Social Media. In *Proceedings of the NSF Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation*, 2007.

[JSFT07] A. Java, X. Song, T. Finin, and B. Tseng. Why We Twitter: Understanding Microblogging Usage and Communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM New York, NY, USA, 2007.

[MG06] G. Mishne and N. Glance. Leave a Reply: An Analysis of Weblog Comments. In *Proceedings of the 3rd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 15th World Wid Web Conference*, 2006.

[MO06] C. Macdonald and I. Ounis. The TREC Blogs06 Collection: Creating and Analysing a Blog Test Collection. *Department of Computer Science, University of Glasgow Tech Report TR-2006-224*, 2006.

[Pin08] José Pedro Pinto. Detection Methods for Blog Trends. Master's thesis, Faculdade de Engenharia da Universidade do Porto, July 2008.

[QRSA07] V. Qazvinian, A. Rasoulian, M. Shafiei, and J. Adibi. A Large-Scale Study on Persian Weblogs. In *Proc. of Workshop on Text-Mining and Link-Analysis*, 2007.

[Sif06] David Sifry. Sifry's Alerts: State of the Blogosphere, August 2006, August 2006. Available in http://www.sifry.com/alerts/archives/000436.html.

[Sif07] David Sifry. Sifry's Alerts: The State of the Live Web, April 2007, April 2007. Available in http://www.sifry.com/alerts/archives/000493.html.

[Sif08] David Sifry. Technorati: State of the Blogosphere 2008, August 2008. Available in http://technorati.com/blogging/state-of-the-blogosphere/.

[TRM03] F. Tricas, V. Ruiz, and J.J. Merelo. Do We Live In a Small World? Measuring the Spanish-Speaking Blogosphere. In *Proceedings of BlogTalk A European Conference on Weblogs*, pages 158–171, 2003.