FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

FEUP

Detection Methods for Blog Trends

José Pedro Gaiolas de Sousa Pinto

Report of Dissertation

Master in Informatics and Computing Engineering

Supervisor: Maria Cristina de Carvalho Alves Ribeiro

Supervisor: Sérgio Sobral Nunes

2008, July

Detection Methods for Blog Trends

José Pedro Gaiolas de Sousa Pinto

Report of Dissertation
Master in Informatics and Computing Engineering

Approved in oral examination by the committee:

Chair: Gabriel de Sousa Torcato David

External Examiner: Pavel Pereira Calado
Internal Examiner: Maria Cristina de Carvalho Alves Ribeiro

16^{th} July, 2008

Abstract

Following the popularity of the web and the large number of documents which are available in electronic form, the search for knowledge in online text collections has been getting bigger. For the last few years, blogs and the specific characteristics of the Blogspace have been in the center of a new research interest by the Information Retrieval community, emerging as an important area for knowledge discovery and becoming a good source for trend analysis in fields such as Business Intelligence and Marketing Research. Topic Detection and Tracking (TDT) has long been studied for timelined corpora, focusing on tasks such as identifying topics in a stream of news articles, locating the first reference to an emerging story, and classifying articles by topic. Being a timelined and extremely hyperlinked domain, the Blogspace is currently in the center of a new wave of TDT applications, with much work focusing on Blog Trend Detection. Even big web search engines are adapting their services to the Blogspace. From a collaboration with the portuguese Internet Service Provider SAPO, we were handed a dataset with more than 50,000 blogs and over 3 million posts to be used as a scholar research resource. To know what we could expect from this corpus while using it for trend detection, our first step was to characterize the dataset based on existing analysis. Focusing on social, evolutive and textual content, this analysis helps to characterize the evolution of both our collection and portuguese bloggers' habits. In this work, we tried to devise new methodologies for TDT that would perform adequately in our corpus. After selecting one algorithm already employed for extraction of important topics in a selected period, we devised variants that showed consistent results when applied to our corpus. We present a comprehensive study on blog research, as well as a simple characterization of the portuguese Blogspace. We also present and explore two tools to retrieve daily hot terms and important monthly topics. They are successfully applied to the entire time span of our blog collection, showing the hype about different thematics in the portuguese Blogspace over the last 5 years.

Resumo

Impulsionada pela popularidade da web e pelo grande número de documentos que se encontra disponível em formato electrónico, a pesquisa em colecções de texto *online* tem crescido consideravelmente. Nos últimos anos, os blogues e as características específicas da blogosfera têm estado no centro de um novo interesse por parte da comunidade científica. Desde a última década, os blogues emergiram como uma área importante para a descoberta de conhecimento, tornando-se numa excelente fonte para a análise de tendências em áreas como a Inteligência Empresarial ou a Pesquisa de Mercado. A Detecção e Seguimento de Tópicos (DST) há muito que é estudada para colecções datadas, com tarefas como a identificação de tópicos num fluxo de notícias, localização da primeira referência a uma história emergente ou a classificação de artigos por tópicos. Sendo um domínio temporal e com várias interligações, a blogosfera está de momento no centro de uma nova onda de aplicações para DST, com muito do trabalho a ser focado na detecção de tendências em blogues. Até os grandes motores de pesquisa começaram a adaptar os seus servicos à blogosfera. De uma colaboração com a empresa fornecedora de acesso à Internet portuguesa SAPO, foi-nos cedida uma colecção com um total de 54.149 blogues com mais de 3 milhões de entradas para ser usada como recurso de investigação. Para saber o que podemos esperar desta colecção de blogues portugueses quando a utilizarmos para detecção de tendências, o nosso primeiro passo foi caracterizar a colecção baseando-nos em análises realizadas por outros investigadores. Focando a análise em conteúdos sociais, evolutivos e textuais, caracterizamos a evolução do nosso conjunto de dados e dos hábitos dos bloguistas portugueses. Por fim, depois de seleccionar um algoritmo conhecido para a extracção de tópicos importantes num período seleccionado, melhoramo-lo com algumas alterações que demonstraram resultados consistentes quando aplicadas à nossa colecção. Apresentamos neste texto um estudo alargado sobre a investigação científica em blogues bem como uma caracterização simples da blogosfera portuguesa. Também apresentamos e exploramos duas ferramentas para devolver temas quentes diários e as tendências mensais mais importantes, mostrando os diferentes temas que estiveram em foco ao longo de 5 anos de Blogosfera nacional.

Acknowledgements

For all the input given to this work since February, when we first met, a word of acknowledgment must be left to both Cristina Ribeiro and Sérgio Nunes. For making the blog collection used in this work available to my research, I must thank João Pedro Gonçalves and all the team behind SAPO.

José Pedro Pinto

"Lucy in the sky with diamonds
Lucy in the sky with diamonds
Ah.... Ah...."

The Beatles

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introduction

It is a fact that people became so familiar with the web that thinking of life without it can be a hard task for many of us. Catching up with the news, connecting with friends or simply looking up for some long forgotten recipe, individuals found out how the web can be useful in their lives. With an estimated size of 55 billion pages increasing every day [1], the World Wide Web is the most valuable information resource available nowadays to the common user. As more and more people discover the web and its possibilities, companies, organizations and individuals are assuring their presence by investing large amounts of money in it with prospects of an even bigger return. However, in recent years a new trend took its place as a web and media authority.

Evolving from a small sporadic hobby to a million dollar job, *blogging* became one of the most enjoyable and even profitable services available on the web. The mainstream adoption of *blogging* by individuals and organizations is still changing the environment of information diffusion since it made it easier than ever to access all sorts of knowledge.

1.1 Blogspace: Reasons for Trend Research

Although Information Retrieval (IR) has proven effective in the identification of authorities on the web, common methodologies are not satisfactory for blogs [2]. Due to it's timeliness and extremely dynamic network, blogs are growing into a separate field of research. But a lot of different areas can be identified as possible paths to follow as a scientific research work. Although the first studies on blogs focused on effects of bloggers age and gender on the Blogspace, researchers nowadays are turning towards extracting communities, topics, opinions and sentiments.

Bloggers' tendency to post entries related to specific topics that change over time can be very useful for social and technological studies. The possibility of self-publishing enables authority bloggers to shape the political ideas and cultural feelings of their readers, that will then react and transmit the topic to readers of their own blogs. This is what makes the Blogspace such a fascinating subject. Besides defining what issues are interesting for some communities, cohesive discussions will be indicators of the pulse of a part of today's society.

Also, the definition of bloggers interests and monitoring of public opinion has a great research potential towards, for instance, business intelligence and marketing. The introduction of a new product in a market can trigger an intensive discussion in the Blogspace, allowing researchers to track, understand, and even predict costumers interests in the given product. Businesses will therefore have better information about their products in the market and the effectiveness and health of the company's image. People will be able to observe the flows and changes of thematic discussions across blogs and sometimes even influence or create them; researchers will detect and predict emerging trends; analysts will have new data available to detect resonance in the community of events triggered around the world.

1.2 Thesis Proposal

This work is based on a blog corpus collected by the portuguese Internet Service Provider (ISP) SAPO[1]. The data has been dumped from their blog search engine database, with 54.149 blogs and over 3 million posts assured to have been mostly written in Portuguese. At the same time, we know about the growing research being done with blogs, as well as the existence of methodologies for detecting trends in blog collections. However, until today we do not know of any application of Blog Trend Detection methodologies in the portuguese Blogspace, in order to detect past trends and today's hot topics. Therefore, we intend to show that:

> It is possible to adapt existing methodologies employed in previous Blog Trend Detection research to perform appropriately on a complex portuguese blog corpus.

To prove this, we first need to deepen our study about existing methodologies and select the most promising one(s). After an overview of the technical and technological requirements to be used in our research for both devising new methods and to develop trend detection applications, an analysis of our blog collection should be made, obtaining

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

features that may be useful when studying trend detection methodologies for our corpus. We will then be able to choose and adapt a solid methodology to use in our trend detection mechanism applied to the portuguese Blogspace. In the end, a tool to automatically retrieve trends in this Blogspace should be presented along with the most important results retrieved by our mechanism.

1.3 Document Organization

Chapter 2 presents a large overview about what exactly is a blog and its potential as a research field. It introduces blogs and the Blogspace, the main terminology used and the different blog-related research categories. Chapter 3 does a characterization of our own collection properties, with a main focus on social behavior, evolution and content, followed by an overview on related works of creation and analysis of datasets. Chapter 4 explains in detail the major part of our work, explaining the application of Blog Trend Detection mechanisms to our collection. After presenting the Frequency Segments methodology used by us, it will introduce topics such as the preparation of a large blog collection for trend detection proposes, the retrieval of hot daily terms and, in more detail, the detection of monthly important topics when applied to our blog corpus. Finally, Chapter 5 summarizes our contributions and provides guidelines for possible future work.

Capítulo 2

Background: Blogs as a Research Field

While blogging is a relatively recent phenomenon, research efforts about blogs have been present since the first blogs started to appear, building on the experiences of various adjacent scientific areas including data mining, social network analysis, game theory and economic research. With the growth of the Blogspace from a few tightly interconnected bloggers to a growing community of millions of users, dedicated means to search, explore, and analyze the Blogspace became increasingly popular among random blog users and scientific researchers alike.

The first part of this chapter gives a large overview about blogs, introducing the Blogspace, as well as the main terminology used when talking about blogs. In the second part, the main research focusing on blogs is presented, with an overview of the main blog-related research categories.

2.1 What is a Blog?

It all started in 1996, with early hand-edited collections of pages containing several dated entries. In 1999, blogs got their first boost with easy-publishing tools like Blogger[1], but only in 2001, with the rise of several discussions around September 11 and the U.S. invasion of Afghanistan, blogs started getting the public's eyes attention, becoming prevalent [3].

Built around the idea of dated entries, or posts, presented in a reverse chronological sequence, blogs allow users (bloggers) to write content focused on observations or discussions ranging from mainstream news to extremely personal behavior. Blogs vary from the work of a diarist or artist/writer, to the self-declared expert and news filter. They can be crafted by a single person or in a collaborative way. Topics are usually centered in

[1] http://www.blogger.com/

the blogger's interests, but they can also be place for a discussion between two or more bloggers with distinct opinions on the same subject. Structurally, a blog also supports comments from readers, links to other blogs in what is called a blogroll and references to own posts in other blogs, or trackbacks.

Blogs are seen nowadays as the best example on how each person can use the Web to make their own personal media. For many, this vast human network is a key tool for obtaining useful information on a daily basis, and with the help of blog clusters like Technorati[2] or BlogPulse[3], users can easily read informal personal opinions about the latest topics. Being a young and dynamic media genre with increasing influence and capability to cultural change, blogs are becoming a significant component of the global information and communication infrastructure.

Trends on the Blogspace usually emerge by outside news, that rapidly adapt themselves to inside discussions. Such emerging trends are seen as topical areas that grow in size and variety at an increasing rate over time. Within a community of interacting bloggers, a given topic may become the subject of intense debate and then slowly fade away, as new topics arise. It is precisely the life cycle of a topic, starting from incipient trend that creates a sticky topic and finishes as yesterday news, that can be used as the object of research.

2.2 Terminology of Blogs

Although blogs may have a variety of forms, most of them appear to have basic features that allow us to call them a blog. The following statements were inspired by Mishne [4], which does a very complete work on both introducing blogs and describing their terminology.

Blog Post A blog is built upon a collection of individual entries, or articles, called posts. Each blog post typically is organized as follows: a title with a succinct definition of the post content, the date of its creation, and its content, or body; most of the times, the author's - also called the blogger - name or another signature is provided too. The most common form of displaying blog posts is by presenting the most recently added ones on the blog's main page, sorted in reverse chronological order.

Blogspace Blogspace is a term encompassing all blogs and their connections, and captures the notion of blogs as a social network. One of the reasons why blogs are so widely

[2] http://technorati.com/
[3] http://www.blogpulse.com/

accepted and studied, is their hyperlinked nature. A blog lives not only of each one of the individuals that read it but specially thanks to all blogs that link to it, opening the blog to the community. Despite being typically called "blogosphere", a term coined in the late 1990s, we will in this work use the term Blogspace each time we endorse the blog domain, since this is more commonly used in scholarly works.

Comments Most blogging platforms like Blogspot[4] or SAPO Blogs[5] allow readers to react to a blog post by writing a comment, which is then displayed after the post itself. This communication mechanism is one of the most important attribute distinguishing blog content from plain web content. Like blog posts, comments usually appear in chronological order (but, unlike the posts, the ordering is not reversed - the most recent comment usually is the last to appear). To prevent abuse, some commenting mechanisms require some sort of user authorization, or are moderated by the blogger.

Trackbacks Trackbacks[6] are a family of protocols that are used to notify a web site about a page linking to it; the trackback protocol is the most common of them, and is often used to refer to the technique as a whole. The main usage of trackbacks in the Blogspace is linking related blog posts. A blogger referencing a post in another blog can use the trackback mechanisms to notify the referenced blog; the referenced post will then usually include a short excerpt from the reference in the original blogs and link back to it. These excerpts, also referred to as trackbacks, typically appear at the end of the post, with the comments to the post.

Web Feeds and Syndication Feeds are documents consisting of structured, usually timestamped, items; they are used to distribute content in a format which is easy for computers to parse. Feeds are almost always XML-formatted; two specific XML[7] formats called RSS[8] (RDF Site Summary or Really Simple Syndication) and Atom[9] are currently used for the vast majority of feeds. Syndication, in the context of blogs, is the process of distributing blog content in standardized format through feeds. Most blog authoring tools support various syndication options. Blog readers can choose whether they access the blog through its HTML interface, or read the contents of the web feed. In the latter case, readers use specialized software to view the feed. These are called aggregators. Typically, readers of a blog who choose to view its content through its web feed subscribe to

[4] http://www.blogspot.com/
[5] http://blogs.sapo.pt/
[6] http://www.sixapart.com/pronet/docs/trackback_spec/
[7] http://www.w3.org/TR/xml11/
[8] http://www.rssboard.org/rss-specification
[9] http://tools.ietf.org/html/rfc4287

the blog's feed, meaning that their software regularly checks for feed updates. The usefulness of syndication and its popularity with web users led online newspapers and other dynamic web sited to adopt it and it is currently prevalent in many non-blog pages.

Permalink Short for permanent links, permalinks are URL's referring to a specific blog post. The pages reached by these URLs are supposedly guaranteed to be permanent. Permalinks first appeared in 2000, as the amount of blog content increased, and a mechanism for archiving older content was required; today, permanent links are a standard feature in almost every blog. Sometimes, a user can only have the full disclosure of a post body once it has jumped to its permalink page, while in the main page only the first paragraphs or a summary of the entire content is shown.

Tags Tags are labels assigned to posts that were created to facilitate organization and navigation of the blog's content by topic. Most blogging tools nowadays allow tagging in various forms, although its use is still a bit chaotic, since one of the characteristics of tags is not having any restrictions about what kind of content a user can include while tagging a post. This mechanism is used both as a way to describe an entry and as an organizational aid for the blog content.

Blogrolls A blogroll can be best described as a list of blogs added by a blogger to his blog main page. These blogs are usually the ones that the author finds interesting or that he follows more regularly. This is seen both as a navigational aid for visitors to the blog, helping them find other related or interesting blogs, as well as a measure of confidence from one blogger to another, acknowledging most of the times a relation between the two blogs.

Archive Being a timeline-oriented medium, most blogs offer a mechanism for browsing the blog content by date: an archive, usually arranged by months or weeks. Besides being useful for the random user, archives can be very interesting for researchers that want to find a simple way to crawl an entire blog trying to fetch every entry online, contrary to crawling web feeds where only a group of most recent posts are shown.

Splogs Spam blogs are generated with two often overlapping goal. The first is the creation of fake blogs, containing gibberish or hijacked content from other blogs and news sources with the sole purpose of hosting profitable context based advertisement. The second, and better understood form, is to create false blogs, that result in a link farm intended to manipulate the ranking of affiliated sites. The urgency in culling out splogs has become all the more important in the last year, evident from the frequent discussions and reports on the issue.

2.3 Research in Blogs

After introducing blogs and the Blogspace, we are now able to better understand the many possibilities researchers face when deciding to work with blogs. We will take a look at blogs not as a tool that enables us to express ourselves but as a domain open to knowledge extraction, discussing the different branches open to research. Next, after presenting the two foundation works of blog mining, we will presented the 5 main areas of analysis where blogs are playing a major part. This will finally lead us to Section 4.1 where the main methodology adapted by us from the literature is thoroughly explained.

2.3.1 Foundations of Blog Mining

Kumar et al. [5] were the first to study in 2003 the evolution of Blogspace, proposing two new tools to address the evolution of hyperlinked corpora. First, they defined time graphs to extend the traditional notion of an evolving directed graph, capturing link creation as a point phenomenon in time. Second, they developed definitions and algorithms for time-dense community tracking, to crystallize the notion of community evolution. They developed these tools in the context of blogs and proved that Blogspace underwent a transition behavior around the end of 2001. Also, Kumar et al. discovered dense periods of "bursty" intra-community link creation, meaning that over some periods within a community of bloggers, a given topic may become the subject of intense debate, and then fade away. At those times, bloggers tend to link to each other more than usual, creating bursts of connectivity and information sharing between them. Bursts could then occur due to increasing activity by one or two bloggers during the time period, while other bursts are the result of many members of the community contributing new links to each other. This was the base study for most Blogspace research being done nowadays.

One year later, in the 2004 World Wide Web Conference, Gruhl et al. [6] presented their studies about the dynamics of information propagation. This is also one of the most cited works, since they were the ones formalizing the notion of long-running "chatter" topics consisting of several "spike" topics generated by outside events or, more rarely, by resonances within the community. They compared information propagation to epidemic diseases and also indicated four significant differing roles played by individuals in the life cycle of a topic. After these two major works, research in blogs became extremely active.

2.3.2 Blog Search and Ranking

The search paradigm became the main mechanism through which web information is accessed since its boom in the late 1990s. By the time blogs emerged, search technology was ubiquitous, and as the Blogspace gained momentum, search engines dedicated to it

quickly came to light. Early platforms such as Blogdex and Daypop (2001) soon were followed by commercial services supporting search: cases of Technorati (2002), BlogPulse and Feedster (2003), PubSub (2004), and Sphere (2005). After 2005, when blogs finally became mainstream, it was the time of the biggest search engines to offer blog search as a separate search medium (cases of Google and Ask.com) or as integrated search results in separate listings (Yahoo).

In most cases, there are two ranking approaches offered by blog search engines: either recency-based, where the latest posts are displayed first, or a traditional combination of keyword relevance and authority, estimated by link indegree. Originally, all search engines focused their retrieval on simple blog entries rather than entire blogs, assuming this should be the information unit in which the searcher has interest. However, Technorati later added their exploration mechanism, where blogs are retrieved, promptly being followed by Google's Related Blogs feature.

Blog search engines distinguish themselves from regular web search engines, with their typical simple user interface with a focus on keyword-based search and feature since their early stages advanced navigational tools. These were designed to take advantage of properties of the Blogspace like its interlinked structure, timeliness and content type. Regarding structural tools, in BlogPulse a user can follow conversations between bloggers, explore hyperlinks from a blog in Technorati and see lists of top-linked posts, blogs and stories in all blog search engines. Timeline-related tools include displaying daily occurrences of search terms over time in the Blogspace, enabling the user to quickly identify bursts in topics. And at last, users can also mine phrases and named entities as well as identify related news stories and books or search for blog tags. Those were some of the content-related mechanism offered by blog search engines.

The challenges that blog search engines face are considerably different from those that web search engines experiment. For one, while the size of data in the Blogspace is much smaller than that in the web, its refresh rate should be of higher importance since usually returning the latest result is crucial. Additionally, contrary to web searches that often focus on early precision, since users only examine the top results of a query, in the blog search recall is equally important, as users expect complete coverage of the keywords they searched for since often they are no more than tracking references to names of products or people. Similar to BlogPulse and Technorati, Joshi and Belsare [7] presented BlogHarvest, a blog mining and search framework that extracts the interests of a blogger, finding and recommending blogs with similar topics and providing blog-oriented search functionalities. It used classification, links and topics similarity-based clustering, and Part of Speech (POS) tagging to provide these features.

Now, for the random web page, the use of link-analysis methods to estimate its authority has had a dramatic effect on the quality of search. Since the Blogspace is an extremely hyperlinked environment, there is a number of link-based methods for authority ranking that have been explored by researchers. Fujimura et al. [8] propose an algorithm called EigenRumor, a HITS-based [9] method that scores each blog entry by weighting the hub and authority scores of the bloggers based on eigenvector calculations. This algorithm assigns a higher score to the blog entries submitted by a good blogger but not yet linked to by any other blogs based on acceptance of the blogger's prior work. Wu and Tseng [10] developed another variation on the HITS algorithm, but this time the link graph is built between blog posts rather than blog pages. Moreover, each link between posts is weighted according to the similarity between the posts and their temporal distance. Hub and authority scores are calculated per post and then propagated to the blog level using several methods, obtaining an average authority of posts in the blog and an overall authority. In another direction, Adar et al. [11] propose a ranking algorithm aimed at identifying the source of the information in an epidemic-like propagation. For this, they develop the iRank, that operates by constructing an information-flow graph of blogs and calculating PageRank [12] values over it. Another version of the PageRank algorithm applied to blogs is the BlogRank [13], using both link graphs and similarity scores and building an enhanced and weighted graph of blogs. With a different direction, Nakajima et al. [14] propose a method to discover bloggers that take an important role in blog threads. Dividing bloggers into two categories - agitators who stimulate discussion and summarizers who provide summaries of the discussion - they are then more likely to identify important blogs with hot conversations. And finally, Ulicny et al. [2] presented last year new blog metrics and improve information retrieval in blogs: relevance, specificity, credibility and timeliness of blog entries, refactoring some metrics commonly used in web retrieval.

2.3.3 Sentiment Analysis in Blogs

Sentiment analysis is a field in computational linguistics that aims to identify, extract and classify opinions, sentiments and emotions expressed in natural language. Most of the work in this area focuses on how to classify the polarity of opinions expressed in texts. Usually, work outside the Blogspace uses data from product reviews, news corpora or message boards, as well as other sources of information. Applications are tools for tracking attitudes of writers in online texts with respect to products or political issues.

Since the Blogspace provides researchers with an unique perspective to people's personal feelings and experiences, research in this field focusing on blogs constitutes one

of the biggest shares on blog mining, with a great part of this work consisting on applying existing methods for sentiment classification to blogs. Chesley et. al [15] use a combination of textual features with verb and adjective polarity to classify sentiment in English blogs, resulting in accuracy levels approaching those reported in domains outside the Blogspace. However, both Ku et al. [16] and Mullen and Malouf [17] report low accuracy in classifying sentiments on blog articles.

Finally, remarking how mood classification in blogs posts is a complex task for both humans and machines, Mishne [18] has given over the last three years several important contributions towards mood analysis. In [19], Mishne shows that in the context of movies there is a good correlation between citations to movies both before and after their debut in theaters and their success in the box office, arguing that sentiment might be effectively used in predictive models when used in conjunction with additional exterior factors. Mishne [20] also describes three simple heuristics that improve opinion retrieval effectiveness by using blog-specific properties like timestamps, comment amount and query-specific spam filters. In other work [21], he experiments with different ways to rank the validity of the opinions presented in each post.

2.3.4 Blog Spam Detection

Ranking high in the results of the big search engines is crucial to web sites, particularly commercial-oriented ones; this led to the creation of the multi-million industry of Search Engine Optimization (SEO), that deals with improving the rank of web pages in search engine results. But while some SEO mechanisms are considered to be legitimate by search engines, some are viewed as search engine spam, aimed at pushing web pages higher than their expected ranking in search results by misleading the algorithm that scores that page.

In the Blogspace, two types of spam methods can be found: comment spam and spam blogs. Comment spam exploit the ability given to visitors to create content which is displayed side-by-side with the post content, to create comment content with links to their own pages so their link-based score can be increased. A variant of comment spam can occur as trackback spam. On the other hand, spam blogs (or splogs) are blogs whose content is automatically created or stolen from more important sources. They are usually used to host advertisements that will benefit from the automatically generated content.

Java et al. [22] proved that spam had a tremendous effect in various blog-related tasks, specially when using spam-vulnerable methods such as HITS. Following that, there have been numerous approaches trying to prevent or reduce spam in blogs. Kolari et al. [23] detect splogs with accuracy levels of 90% by using a machine learning approach; the work

follows a philosophy of blog level spam detection instead of focusing on single posts, using the fact that splogs usually differ from legitimate blogs in terms of language and link structure. A similar approach is used by Narisawa et al. [24], regarding detection of the main differences between legitimate and spam blog languages. Finally, Lin et al. [23] go a step further by using not only language and link proprieties but also structural and temporal properties that also differ in spam blogs from typical blogs.

However, if one wants to avoid spam blogs in their collection while crawling the web, it is better to block splogs prior to fetching them, basing this in their ping notifications to blog ping servers. Hurst [25] shows some differences on ping patterns of both type of blogs, showing evidences that these same differences can be used to make legitimate blogs and splogs apart from each other. At last, although research on comment spam is not abundant, Han et al. [26] also proposed an approach where bloggers would manually filter spam comments in their blogs, propagating their decisions to trusted bloggers.

2.3.5 Blogs as Consumer Generated Media

Consumer Generated Media (CGM) refers to experiences and opinions produced by consumers of products, brands, companies or services. Being a particular type of user-generated content, CGM can be found in contexts such as emails, discussion boards, product review sites and, of course, blogs. The Blogspace is of special interest as a source of CGM for a few reasons: for once, the informal, unedited content of blogs is considered by most marketers and consumers as a more raw reflection of the consumer attitude than other sources; second, the popularity of some blogs exceeds that of some more technical boards, which enhances high-influence bloggers with an enormous power as opinion makers; finally, since bloggers are heavy users of technology, as well as being usually the first to adopt it, they serve as an important focus group for analysts.

With the increasing volume of blogs and other public online boards, commercial companies that target their business on mining business intelligence from these sources emerged. The best example of this trend is Nielsen BuzzMetrics. The mechanisms these companies use are aimed at discovering what the consumer thinks about a given product or company, by examining CGM. The methodology used by Nielsen Buzzmetrics [27], for instance, consists of a platform that does its own crawling and information retrieval as well as sentiment and social network analysis; a similar platform is discussed by Tong and Snuffin [28]. A demonstration of the potential of blog analysis in the field of marketing was described by Gruhl et al. [29], observing that bursts in book sales are, in many cases, preceded by similar spikes in references to these books in the Blogspace. By comparing bursts of blog entries that refer to a given book which boosted sales according to Amazon,

they imply that a sudden increase in blog mentions is a potential predictor of a spike in sales rank, affirming that by tracking references to books in the Blogspace we can gain the ability to predict future sales boosts. Also worth mentioning is the work of Aschenbrenner and Miksch [30] who have done an exhaustive presentation on the past, present and future of blog and all the different possibilities of adapting CGM inside corporate environments.

2.3.6 Blog Trend Detection

One of the most important branches of Blog Mining is the identification and tracking of topics inside the Blogspace. Topic Detection and Tracking (TDT) has for long been studied for timelined news corpora, with tasks such as identifying topics in a stream of news articles, finding the first reference to an emerging trend or classifying articles by topic. The Blogspace, being a timelined and extremely hyperlinked domain, the Blogspace proves to be an especially interesting domain for TDT applications.

Simple methods for trend detection in the Blogspace have appeared, the most promising ones being proposed by Hurst [31] and Oka et al. [32]. The first proposes a number of possible trend models (linearly increasing, bursty, etc.) and shows that by calculating the similarity between models and occurrences of terms over time in blogs, different trends can be uncovered. Oka et al. focus their work in the extraction of topics by following the frequencies of terms over time and measuring a sum of deviations; terms that are considered important are then matched with co-occurring terms to form more descriptive topics. Similar work has also been done matching popular terms in the Blogspace with popular terms in mass media, where researchers try to identify overused terms in a given time-slot [33, 34]. The methodology employed by Oka et al. will be detailed in Chapter 4, since this was used as foundation for our work.

More complex methodologies are used by Glance et al. [35] in an application that combines data mining, information extraction and Natural Language Processing (NLP) algorithms for discovering trends across a subset of blogs. They found trends through most-popular links and use of phrase and name finding with the aid of English language frameworks. The result of this work was the now famous BlogPulse website[10]. Taking another direction, Teng and Chen [36] use textual, temporal and interactive features to achieve better results in the detection of bloggers' interests in given topics. Studying term, post and comment frequency, they are able to identify the main interests of each individual. Finally, also worth mention is the work of Qamra et al. [37] who use a Content-Community-Time model to extract discussions, or stories, from the Blogspace.

[10] http://www.blogpulse.com/

They use hyperlinks between blogs to first create a community structure that is then used to group different terms in the same story.

Chi et al. [38] improve on trend-analysis methods based on term counts by combining the temporal profile of a blog with its link structure; the results are more robust trends as well as the identification of signs in keyword usage patterns which are not usually easily observed. In a similar way, Zhou et al. [39] incorporate a social network analysis tool with a trend-detection approach, enabling the identification of trends using interactions inside the community and locating, for each trend, the most influential actors. But besides content analysis based methods, also links can be used to track topics in blogs. Wu and Tseng [40] employ a HITS-based method for identifying hot topics in the Blogspace. More recently, two works were published with very satisfactory results for the Trend Detection subject [41, 42]. In the first, Gamon et al. present a framework for detection and tracking of articles from news feeds and tagging them with blogs that cite them. Nallapati and Cohen employ a combination of the Probabilistic Latent Semantic Analysis (PLSA) and the Latent Dirichlet Allocation (LDA) - models that exploit co-occurrence patterns of words in documents to discover clusters of words, or topics - into a single framework, being able to visualize topics as well as the most influential blogs on each thread.

2.4 Summary and Conclusions

A Blog is a website maintained by one or more individuals, composed by a sequence of entries usually presented in a reverse chronological sequence. Topics are often centered in the blogger's interest and can also contain opinions expressed by readers in the form of comments.

The Blogspace, name given to the set of all blogs and the connections between them, is nowadays seen as a form of social media used by many as a key tool for obtaining useful information on a daily basis. News that appear in the media in the morning, soon will be talked in the Blogspace, thus creating discussions that can be seen as emerging trends of conversation.

To successfully employ mechanisms that automatically detect topics in blog's conversations is precisely one of the main objectives in today's blog research, but scientific research in the Blogspace can be roughly divided in 5 fields:

- Blog Search and Ranking aims to improve results of blog search engines. By distinguishing themselves from traditional web search, blog search engines take advan-

tage of specific properties of the Blogspace like its interlinked structure, timeliness and language. Problems like the refresh rate of the Blogspace or the search recall are addressed by researchers who look for ranking algorithms for blogs.

- Sentiment Analysis in Blogs aims to identify, extract and classify opinions, sentiments and emotions expressed in natural language. Since the Blogspace provides researchers with an unique perspective to people's personal feelings and experiences, research in this field focusing on blogs constitutes one of the largest shares on blog mining with studies about polarity of sentiments being the hot area of the moment.

- Blog Spam research tries to clump two types of spam methods: comment spam and spam blogs. While the former exploits the ability to create comment content with links to pages whose rank a spammer wants to improve, the latter enables the creation of blogs that produce income from advertising by presenting in their body highly ranked terms. A lot of research is done with considerable success on detecting spam, be it by the language used, structural proprieties or even ping notification patterns.

- Blogs as Consumer Generated Media refers experiences and opinions produced by consumers of products, brands, companies or services in their blogs. This field has received great focus by private companies, as it is a very profitable field. The best example of this is given by Nielsen BuzzMetrics, allowing companies to purchase services to get insight from consumer generated media.

- Finally, Blog Trend Detection tries to identify and track topics inside the Blogspace. Either recurring to predefined models, following sequent occurrences, using NLP algorithms or finding communities with the same topical discussions, trend detection in the Blogspace proves to be an especially interesting domain for Topics Detection and Tracking (TDT) applications.

A lot of research has already been done on blogs and the Blogspace over the past few years. However, as far as we know, no one has tried to employ any of these methodologies to the portuguese Blogspace, documenting possible results obtained. The next chapters will try to change this panorama, by introducing two different studies on the portuguese Blogspace: the characterization of a large collection of portuguese blogs, and the detection of blog trends in the portuguese Blogspace over the last year.

Capítulo 3

Analysis of the Dataset

Since hundreds of blogs are created on a daily basis, it is extremely difficult to successfully create a collection that represents the entire network. Over the last five years, SAPO[1] has been crawling feeds of blogs and adding them to their blog search dataset. It is this collection of feeds, containing a total of 54,149 blogs with over 3 million posts that was made available by the portuguese company for research purposes. Although our aim is to use this collection to develop methodologies for topic detection, we will first present in this chapter a simple characterization of the given dataset. It should be noticed that this is a collection of feeds, built by a crawler working daily in an extremely changeable social network and that it may not be representative of the entire portuguese Blogspace.

In the next sections we will do a characterization of the collection, trying to focus on Social Behavior, Evolution and Content; in the end, some of the most relevant work related to this characterization is mentioned, followed by some considerations about the data collected and ideas for future work on the same corpus. The analysis and considerations detailed in this chapter are the result of a collaboration with José Castelo Branco [43].

3.1 Collection Properties

From the initial 54,000 blogs present in the dataset we removed around 4,000 that were outside the time span between January 2003 and December 2007. This left us with 49,940 blogs with 2,933,735 posts spanning the last five years. Each blog is achieved as a sequence of syndicated feeds, whith first posts being the ones contained by the first feed retrieved by the crawler. The rule used to insert a blog in the dataset was simple. Starting from a set of selected blogs, their links are used to detect new ones. The fetcher would

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

store new blogs to the dataset if a linguistic analysis revealed they were written in portuguese from Portugal (pt_PT).

The blogs in the collection can be split in 3 categories according to blog providers. The two big groups are the SAPO hosting, with 52% (25,768 posts) of blogs in it and the Blogspot hosting[2], with 47% (or 23,378 blogs). The remaining 1% is composed by a set of blogs from different hosts, with a total of 794 blogs. However, we can see from Figure 3.7 that the number of posts from Blogspot is larger than the one from SAPO.

3.2 Collection Analysis

To better understand the contents of the collection, we proceed with an analysis focusing on social behavior — trying to explain the posting habits of bloggers in the collection — evolution — depicting the growth of the collection over time — and content — presenting the usage of outgoing and incoming links in blog entries.

3.2.1 Social Analysis

Some social extrapolation can be made from the portuguese blogging habits. Bloggers like to write throughout the day, with small stops during lunch and dinner time. Local maximums can be seen around 12:00 and 16:00 while the more frequent time to post is around 23:00. Figure 3.1 shows the number of posts per hour in the three already mentioned groups of providers. Although they react in the same way, minimums and maximums show up in SAPO one hour prior to Blogspot. Although an one hour lag between the times in SAPO and the ones in Blogspot seems to exist, nothing can be inferred from here since we concluded that this may be due to the process of storing the retrieved feeds into the database. After inspecting and comparing the data available and the feeds still available on the web, we concluded that some blogs have different timestamps and notations in their feeds and in their posts.

[2] http://www.blogger.com/

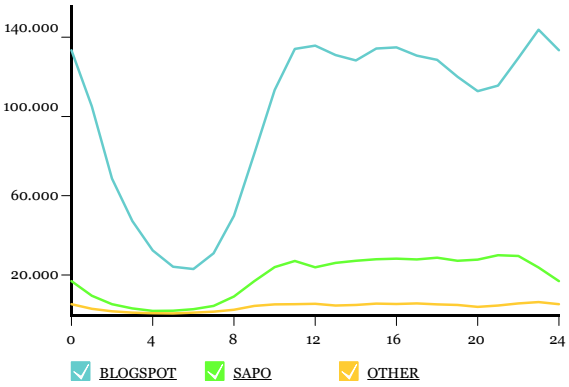

Figura 3.1: Posts by Provider per hour

Figure 3.2 depicts the posting habits by month separated by year. Knowing that the corpus grows every day, it is with no surprises that we see a constant growing in the monthly data. However, even with this behavior we notice slowdowns or even breaks in the Summer and Christmas holidays. The graphic has a big boost between February and April 2006, representing a 300% posting increase. This situation should be studied in the future, since it may be connected to some different algorithm used in the daily crawl, as our next section will again show.

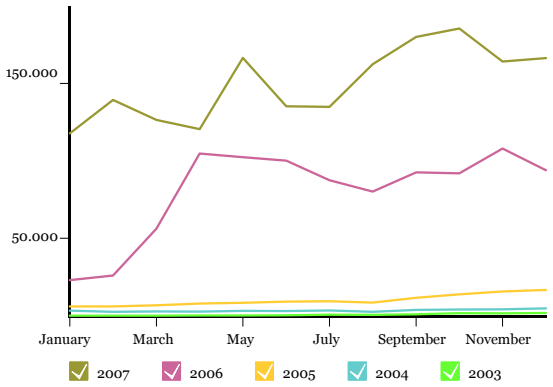

Figura 3.2: Posts per Month over years

Finally, Figure 3.3 shows the blogging habits throughout the week, showing a maximum of activity on Mondays and then a slow decrease until Friday, and a steep fall during the weekend.

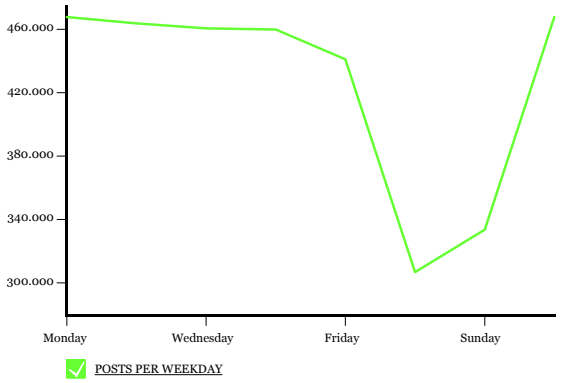

Figura 3.3: Posts per Weekday

3.2.2 Evolution Analysis

By doing an analysis of the corpus size through time, we detected two major changes in the growing behavior of the dataset. Both in April 2006 and February 2007, the growth of the corpus suffered slowdowns, probably due to some shift in the feed crawler. This is depicted in Figures 3.4 and 3.5. It appears, since there is an increase of blogs from SAPO, that the crawler stopped registering new blogs from Blogspot.

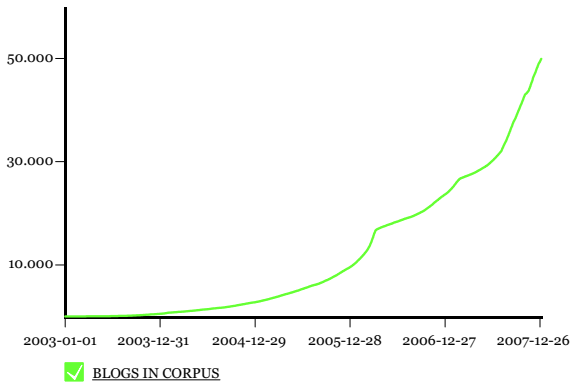

Figura 3.4: Blogs in Corpus over time

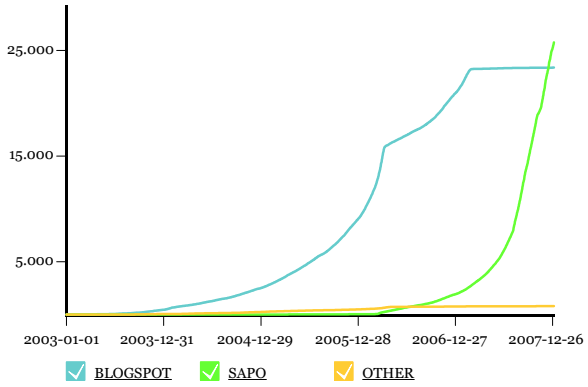

Figura 3.5: Blogs by Provider over time

Figures 3.6 and 3.7 show the number of posts for each day over the 5 year period. In March 2007 a huge drop in the Blogspot posts number is noticed, while the number of posts from SAPO tends to grow. The biggest boom on posting is in April 2006.

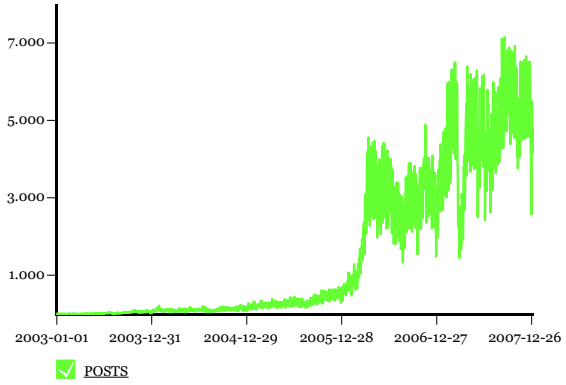

Figura 3.6: Posts over time

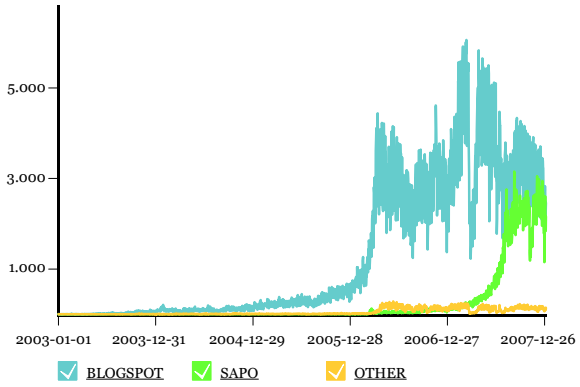

Figura 3.7: Posts by Provider over time

One of the most curious results is the one depicted by Figure 3.8, showing the number of posts per blog. Although there is nothing new about the long tail in this Posts/Blog plot, the local maximum at the 25 posts leeds us to think that something wrong went out with those feeds. This has yet to be studied, since the analysis of those 2,002 blogs was inconclusive. One possible explanation is that for a long time the crawler would just detect the blog, and then stop retrieving more posts as the time went by. Being true, this would be one of the reasons to be sure that this dataset could not be representative of the activity going around the portuguese Blogspace.

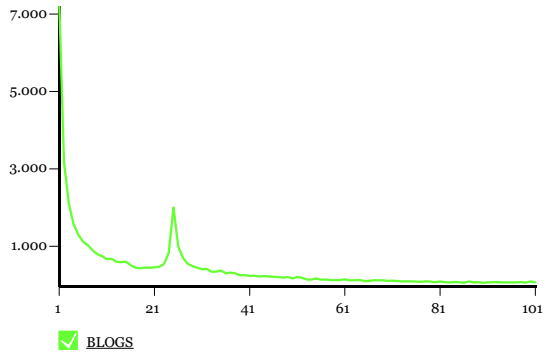

Figura 3.8: Posts per Blogs

3.2.3 Content Analysis

Concerning the content of the posts, some preliminary analysis was done regarding link and word usage. Looking at Figures 3.9 and 3.10 we can say that bloggers in our collection use more self-references than out-links to other community blogs. However, both out-links and in-links are experimenting an interesting growing behavior that can lead to a stronger community in the near future.

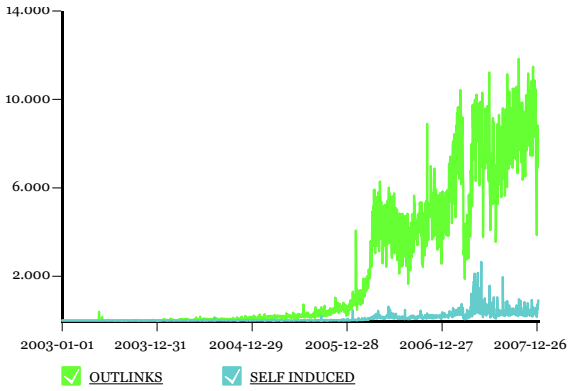

Figura 3.9: Outgoing links over time

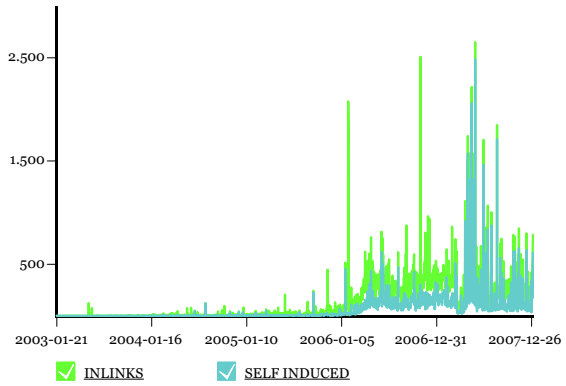

Figura 3.10: Incoming links over time

In what concerns word usage, posts are usually short: half of the corpus is composed by posts with a length ranging between 0 and 75 words. The average length of post is of 160 words.

3.3 Related Work

Some work has already been done regarding the characterization of blog collections. Since the Blogspace has caught several researchers' attention, the need for big amounts of data for experiments has grown. In several projects scientists have been using different corpus, and since the characterization of the dataset to work with is a fairly common step on a project's development, there is already a well built background in which we can support our work.

In 2006, MacDonald and Ounis created the TREC Blogs 06 Collection, and analyzed the data they obtained [44]. The corpus they got took them 4 months to construct, and in that time they included highly popular blogs, regular ones, and also splogs. A crawler was used to check as much feeds as they could per month, and for each feed they would extract all the permalinks. In this way they obtained 100,000 different blogs, with a total of 3 million posts. They concluded in their report that the date information present in the XML feed may be unreliable, as it was observed that a large number of permalinks did not have date information on their feed, and that others presented a wrong date reference, such as before 98. While analyzing the daily habits of a blogger, MacDonald and Ounis observe patterns in the user behavior that generally we can also observe in our own data. They observe that the posting frequency decreases on the Christmas holiday and on weekends, and that the highest posting activity on a day is around 2:00 pm, followed by the lowest, at 3:00 pm.

Qazvinian et al. [45] took up a study on Persian weblogs, in 2007. Using a HTML parser, they gathered their corpus from the largest host for weblogs in that country, the PersianBlog. The data was then converted into XML format, obtaining in the end 80,000 XML files spanning monthly archives of all blogs retrieved. The corpus obtained contained around 22,000 blogs and 348,000 posts, and had one characteristic which neither our dataset nor TREC BLOG06 had: it contained the comments for each post (1,257,561). The main conclusions from this study was that the majority of comments on blogs were made by people who had blogs in the Blogspace; the number of comments by week was distributed quite evenly during the week, except for Thursday and Friday, when there is a decrease in this number; and that the new year holidays and beginning of the academic year caused a lower rate of comments, while the presidential elections increased their numbers significantly.

Another interesting analysis was made available by Kolari et al., in 2006 [46]. With the intention of analyzing spam blogs (splogs) and creating solutions for better avoiding them, they use a dataset made available by BlogPulse spanning 21 days in July 2005 with 1.3 million blogs included, plus their metadata. They show that in-degree and out-degree distributions of splogs do not follow a power-law curve and linked the ping times in public ping servers with the blogs and splogs activities. Comparing the graphic for Italian blogs and English blogs, they conclude that english written Blogspace is much more prone to spam. Through this analysis they were able to conclude that 88% of the pinging URLs are spam blogs.

3.4 Summary and Conclusions

In this chapter, we presented the blog corpus that we will use in the next step to employ trend detection methodologies. The results presented here are still introductory and need more work. Although we can leave some information on how the SAPO crawler works, we still cannot state that the resulting dataset is representative of the portuguese Blogspace. From this exploratory work, a more focused one should take place to analyze more carefully situations like the break in the Blogspot blog count and in the posts from March 2007 on. It would also be interesting to solve the mystery surrounding the local maximum around the 25 post blogs. We also don't know, for instance, how many blogs are still active, nor if the graphical shapes retrieved from the analysis are a result of a blogging behavior or mislead by the spider activity. Therefore, we think that although this chapter draws the lines and directions for a strong case in analyzing this Blogspace, more work needs to be done to mature it and to better characterize the entire collection.

Capítulo 4

Detecting and Tracking Blog Trends

After presenting the main research already done on detecting and tracking blog trends, as well as delivering an overview of the dataset used in our research, we are now able to bring both things together and explain our own work, detecting interesting topics inside our blog corpus over a selected period of time. Our main goal was to build a tool that, faced with a considerable real world corpus, could automatically detect monthly trends — topics that were relevant for a given month — and hot terms — words that are important for a given day.

We will divide this chapter in four sections. Section 4.1 describes the core methodology used in our work, so that the requirements for building a system for automatic trend detection can be understood. Section 4.2 describes how data is handled and prepared to be used by the trend detection mechanisms. These are explain in Sections 4.3 and 4.4, where a method to find daily hot terms and another to detect monthly trends are presented.

4.1 Extracting Topics Through Frequency Segments

The approach employed by Oka et al. [32] and presented in the annual World Wide Web conference in 2006 is used as the core methodology behind our work. Oka et al. extract important topics that appear in weblogs during a defined timespan. We will present this methodology, as well as those that we consider to be the main advantages and disadvantages.

Frequency Segments The reason for choosing this method over others presented in the literature had to do with its simplicity, sustaining itself only in statistical measures, allied with the satisfying results achieved by it. The concept of Frequency Segment is rather easy to understand. Taking a sequence of frequencies of one term presented in a corpus for each day of the defined timespan, a new segment will start for each time this term

frequency is equal to 0. After this, a single value is obtained for the entire segment, as the sum of all the frequencies for the segment, which is then used together with all the other segments in the timespan to calculate this segment deviation. The final result will give the segment weight in the timespan as a score, that can be compared with scores for other terms in the corpus.

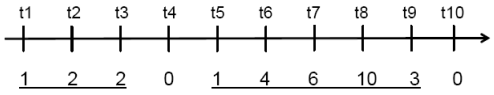

Figura 4.1: Sequence of frequency of occurrences for a term [32]

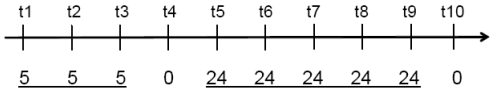

Figura 4.2: Segment sum sequence [32]

Take Figure 4.1 as an example of a sequence V elements, one for each day, where each element D represents the Document Frequency (DF) of a term T in one day, given by the number of documents in which the term T appears on that day. The idea is to return a $score(T)$ to be used to rank T among the other terms in corpus, for a time period spanning $size(V)$ days. This is done by measuring the dynamics of V in units of segment, which are by definition blocks of sequential occurrences of the term T. In this case, Figure 4.1 shows the sequence $V = (1,2,2,0,1,4,6,10,3,0)$, meaning that the term occurred 1 time in the 1st day, 2 times in the 2nd day, 2 times in the 3rd, and so on. Here, the first step while analyzing the sequence of occurrences is to identify two segments: $S' = (1,2,2)$ and $S'' = (1,4,6,10,3)$.

After this, each segment S is weighted with two parameters. The *absolute* strength is the sum of occurrences inside S, called *segmentsum* and denoted as sum_S. The *relative* strength would be the deviation of sum_S when compared to the average of sums for all the other segments for term T, called *segmentdeviation* and denoted as dev_S. Equations 4.1 and 4.2 show how the segment deviation dev_S is calculated; note that σ indicates the standard deviation of the term T on the selected period.

$$z = \frac{sum_S - mean(sum_{S'}, sum_{S''}, \ldots)}{3\sigma} \tag{4.1}$$

$$dev_S = \begin{cases} min(z, 1.0) & \text{if } z \geq 0 \\ max(-1.0, z) & otherwise \end{cases} \tag{4.2}$$

Back to our example, calculating both segment sum and deviation for the segments results in $sum_{(1,2,2)} = 5$, $dev_{(1,2,2)} = -0.24$, $sum_{(1,4,6,10,3)} = 24$, $dev_{(1,4,6,10,3)} = 0.24$; these values are better understood when graphically depicted. Figure 4.3 shows the comparison between the DF values for a term T in a 10-day period and the segment sum for both segments identified, while Figure 4.4 gives a graphical representation of the same segments' deviation.

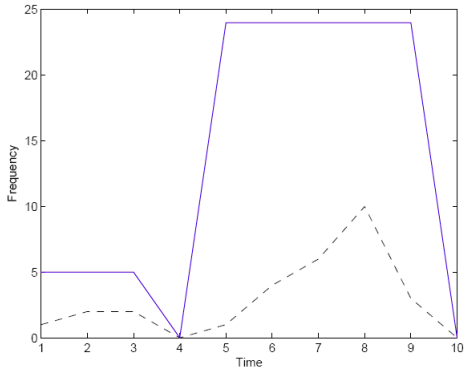

Figura 4.3: Frequency of occurrence and segment sum [32]

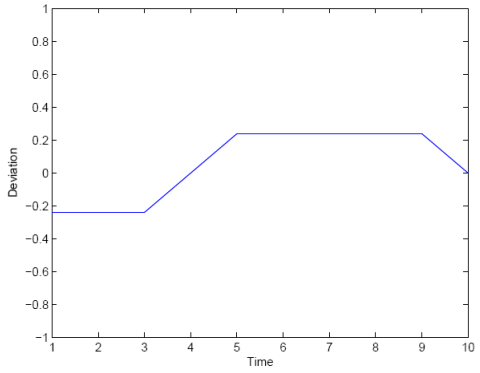

Figura 4.4: Deviations of segment sum [32]

Figure 4.5 plots the deviations and sums of segments for all the terms used by Oka et al. in their experiments. A notable feature is the vertical line along the zero value for segment deviations (in the x-axis). All terms presented in this line are the ones that only had one segment in the sequence of occurrences, meaning that it was a term that occurred everyday during the explored period of time. To obtain interesting terms, researchers

retrieved terms with high values for both deviations and sums, since those were the ones representing greater or sudden increases in occurrences. The 3,200 terms retrieved by them appear in light gray in Figure 4.5.

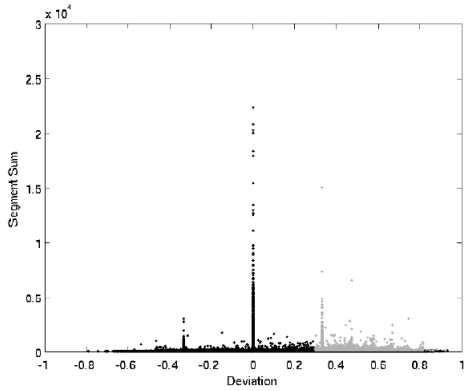

Figura 4.5: Deviations and sums of segments of 41,000 entries [32]

Ranking Terms and Finding Related Terms It has been mentioned before that terms with higher values for segment sum and/or segment deviation are by hypothesis the ones that should be considered with more importance in the selected period of time. Each segment is then given a value $score(S)$, used to rank all segments of a term T; this score is calculated using Equation 4.3 with an empirically determined α, to assign relative weights to the segment sum and the segment deviation; in their experiments, Oka et al. use $\alpha = 0.11$.

$$score(S) = \alpha \times \log(sum_{segment}) + dev_{segment} \tag{4.3}$$

In the end, all terms in the corpus will be ranked by a $score(T)$, corresponding to the maximum value of $score(S)$ for each term. The larger the score, the more relevant the scored term is.

Extraction of Detailed Description of Topics Finally, a more detailed description of each topic can be extracted by relating each term A to all the other terms that co-occurred in the same posting. For each term B that occurred in the same entry as A, the similarity — or correlation — between A and B is given by

$$correlation(A,B) = \frac{< A - mean(A), B - mean(B) >}{\|A - mean(A)\| \times \|B - mean(B)\|} \tag{4.4}$$

Problems of this Methodology After studying the methodology employed by Oka et al., our main conclusion was that although the algorithm looked very strong on retrieving

terms that experimented strong spikes — for being mentioned only for very short periods of time — it would not be able to detect terms that had long chatters running — words that were mentioned almost every day in the data collected, thus consisting in one segment only. While this was not a problem for Oka et al. since they only use a sample of a very large collection (42.000 postings from a collection of more than 8 million of entries), in our case this will be a problem, since we want to detect trends using more than 3 million posts. The next sections will explain how we do retrieve important topics from our data collection, with the Frequency Segments methodology as background.

4.2 Indexing a Collection of Portuguese Blogs

The idea behind creating an index is to optimize speed and performance while finding relevant documents in a search. Suppose we want to find a term in our blog corpus and we don't have built an index. A search engine would have to scan every document in the corpus, which would require a considerable amount of computing power and time. For instance, while our index with around 3,000,000 blog entries can be searched within milliseconds, a sequential scan of every term in 3,000,000 documents might take hours.

Considering our trend detection mechanism, our biggest demand in the short run is to know the daily frequency of occurrences for thousands of terms, and therefore we need a relatively fast way to retrieve these values. Since our corpus is only available in a relational database, we will first have to create a set of files, with the blog entries in our dataset, that can be read by the index creator. Once this task is complete, a list of the words present in the corpus along with their Term Frequency (TF) — the number of times a given term appears in the collection[1] — in the entire corpus is compiled; this list is used to decide which words are considered while automatically searching for topics. Each of the steps will now be explained in more detail.

4.2.1 Creating the File Set

The collection analyzed in Chapter 3 is stored in a database. The first task that needs to be done is to parse the entire dataset to a format that can be both read by the index creator and useful for future research proposes. As in [44], we decided to create a collection of Extensible Markup Language (XML) files.

[1] in IR, the value of TF is usually given to one document instead of the entire collection; this definition is not considered in this document.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<entry
id="43a11468332b000d98bf2fa85e58a950"
basename="baldedegelo.blogspot.com"
url="http://baldedegelo.blogspot.com/88228720.html"
>
<author>
<![CDATA[Marco Aurelio]]>
</author>
<title>
<![CDATA[Oh, meu Deus. Acho que eu devia ter deixado o pood...]]>
</title>
<postdate>
2003-01-30 08:00:05
</postdate>
<body>
<![CDATA[Oh, meu Deus. Acho que eu devia ter deixado o poodle
em casa e mandado minha namorada para um spa ou algo assim.
Seria mais seguro. Foi]]>
</body>
</entry>
```

Figura 4.6: Example of XML file for blog entry

Each one of the blog feeds would be parsed to an XML file similar to the one in Figure 4.6. Each one of the files would be stored in a directory according to the month and year of the post. An identifier is generated for each entry as a Message-Digest Algorithm 5 (MD5) value of the entry's URL, and all the relevant information about the post is saved: blog URL, post URL, author, title, date, and body. The final result is a collection of 2,933,735 XML files divided into 60 directories with a total disk usage of 13.2 GB.

4.2.2 Indexing the Documents

An inverted index is a data structure that stores a mapping of a text collection, to support full text search. When accessed by some search mechanism, the index retrieves a list of references to documents in the collection where the search terms occur, as well as a list of other texts — features of the indexed document — like the name of the indexed file or the position of each word within a document, the latter supporting functionalities like phrase search. In the indexing phase, documents are prepared to be used by an IR system. The raw document collection is prepared into a representation of documents that should be easily accessed; from each of the documents, we get a representation of its text. Transforming a document into an indexed form involves the use of a library or a set of regular expressions, parsers, and eventually a library of stop words and other filters that can later

be added to improve the index response to queries.

The first step when indexing a collection is called Document Linearization. This is the process by which a document is reduced to a stream of terms, and it is usually done in two steps:

1. Markup and Format Removal, where all markup tags and special formatting are removed from the document.

2. Tokenization, in which all remaining text is parsed, lowercased and all punctuation is removed.

After linearization, the documents should be filtered. Here, filtration refers to the process of deciding which terms should be used to represent the documents. Usually, a stop-list of terms to be removed is used, meaning that high frequency terms, or stop-words, are removed since they are usually poor discriminators of a document or topic. The next step is stemming, where terms are reduced to their stems, or root variant. For instance, "computer", "computing" or "compute" would be reduced to "comput". While there is no doubt that stemming ensures that documents containing variations of a given queried term are considered in the final answer set, too much stemming is usually not practical and can annoy users. For example, if stemming is done, the search engine can not show exact matches first because only the stemmed version is stored in the index. The final stage in most IR indexing applications is weighting. Weights are scores given to terms according to some predefined model such as scoring by TF or Inverse Document Frequency (IDF) value — obtained by dividing the number of all documents in a collection by the number of documents containing the term, and then taking the logarithm of that quotient.

In our experiments, we use Apache Lucene[2], a full-featured text search engine library written in Java, used by several web search engines. Indexing all of the fields inside the XML files but without storing its body, so that size would not exceed free space in the server, we end up with an index with 3,882,914 words, taking 3.8 GB, less than 1KB per word. We create a simple indexer that, while using markup removal and tokenization, does not use any kind of stop word removal nor stemming. This is done since the index is used to easily retrieve both frequencies and locations of terms, rather than being used as a typical tool for a web search engine.

[2]http://lucene.apache.org/java/docs/

4.2.3 Retrieving Frequencies of Terms

As explained in Section 4.1, to detect if a term is more important than others to a given time frame we first need to know the frequency of occurrences of that term over the selected period of time. That being said, if we want to create a list of every important term from 2003 to 2007, we need to know the frequencies of occurrences of thousands of terms in this period of time. To avoid retrieving this values while calculating the score associated with a term and a time frame, we should first create another data structure that can easily be accessed when scoring terms.

term	TF	DF
de	18,598,356	2,300,973
para	4,947,337	1,529,522
natal	107,324	59,270
casa	420,945	264,772
menina	50,945	3,5716
iphone	3,179	1,549
mugabe	1,956	1,117
azerbeijão	254	212
ginseng	148	100
powerbook	105	88
procrastinação	100	77
previlegiado	74	69

Tabela 4.1: Example of terms in corpus with their TF and DF

Since our corpus is composed by almost 4 million terms, we define a threshold for the DF of a term above which a term is selected to be part of our trend mechanism. By inspecting the dataset, we decide to remove all the words with $TF < 100$, keeping 105,325 terms to be used by our trend detector system. Table 4.1 shows an example of some of the terms in the corpus together with their TF and DF.

date	DF
2007-12-15	0
2007-12-16	0
2007-12-17	0
2007-12-18	1
2007-12-19	0
2007-12-20	0
2007-12-21	1
2007-12-22	0
2007-12-23	0
2007-12-24	0
2007-12-25	0
2007-12-26	1
2007-12-27	129
2007-12-28	93
2007-12-29	29
2007-12-30	22
2007-12-31	14

Tabela 4.2: Example of listed frequencies for term “benazir”

The last step is the creation of the new file set, containing a file for each of the 105,231 terms, with a listing of the DF of the term for each single day from January 1st 2003 to December 31st 2007. The file is named with the MD5 value of the term, so that it can be easily tracked by the trend detector system. The entire file system occupies 2.5 GB of disk space. Table 4.2 has an example for the file with the list of occurrences for the term “benazir”.

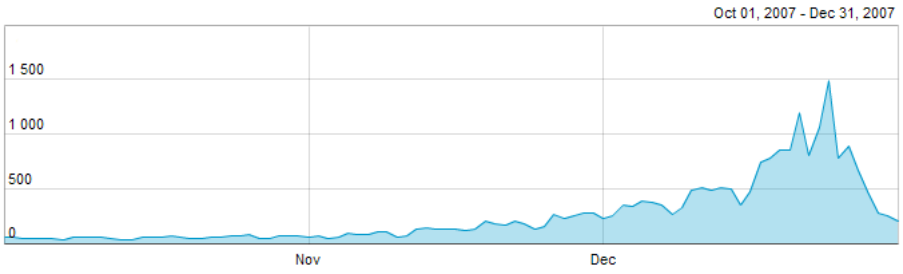

Figura 4.7: Trend for “natal” over last 3 months of 2007

After this step, we are now able to visualize trends for each one of the 105,231 terms parsed, by simply inserting a term’s frequency values in a graphic. Figure 4.7 has the visualization of the trend for the word “natal”, the portuguese word for Christmas, over the last three months of 2007. The next two sections will now explain how to use the file system created to retrieve, for the entire corpus and time span, hot terms for each day and, in more detail, important trends for each of the months.

4.3 Detecting Hot Terms

Similar to what Google does with its trends search engine[3], we employ a mechanism for detection of hot terms, words that are important for one given day - usually, the present day. In our experiments, we try to retrieve hot terms for December 31st 2007, since this was the last day present in our collection. Figure 4.8 shows the frequency of occurrences for terms "ano", "novo" and "2008" during December 2007. The sudden burst of occurrences in the end of the year to these three terms are easily explained for their relation to the New Year's Eve topic. Our desire is to create a simple method that can easily detect bursts like this for one single day.

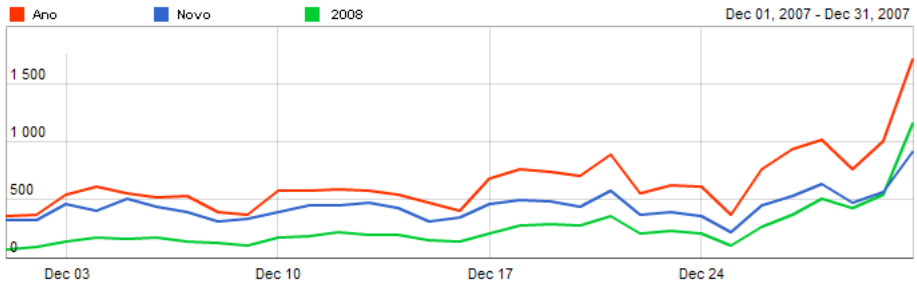

Figura 4.8: Trend for "ano", "novo" and "2008" in December 2007

If the algorithm works as it is expected, it would be interesting to retrieve hot terms for every day in the corpus, so that we could have a retrospective tool like Google Trends has. Hot terms can be seen as finding out the fastest rising terms considering only the given day, thus meaning that the terms retrieved are the ones representative of what was the main discussion for that day or what are the terms that will be the future discussion in the Blogspace for the next days.

For the detection of important terms in one day, we have decided to rank each term T according to its rate of change in frequency of occurrences in the selected day d comparing it with the average number of occurrences for the past 10 days:

$$score(T_d) = N_d - \frac{mean(N_{d-10}, N_{d-9}, ..., N_{d-1})}{N_d} \tag{4.5}$$

If a term sees its popularity rising for one day, this score will be high as long as the term's occurrences over the past 10 days did not expect large bursts that could inflate the value of $score(T_d)$. Remember that we want to retrieve terms that are important for that day, and not for the entire period of 10 days. This simple score will give fastest rising terms for that day a bigger importance compared to all the other terms in the corpus.

[3] http://www.google.com/trends

Although the method is simple, it proved effective when applied to our corpus. Table 4.3 presents the list of results obtained, for December 31st 2007.

2008	ano	realizem	passas
entradas	resoluções	serpa	2007
proibição	amizades	fumadores	concretizem
novo	badaladas	réveillon	amarelas
cadilhe	desejos	silvestre	marisa
entrem	desejo	balanço	saídas
euromilhões	artifício	cálculo	finalistas
31	saúde	discotecas	repleto
termina	fumar	finda	ginástica
montagem	science	realize	passagem

Tabela 4.3: Hot Terms for 31-12-2007

As expected, terms related to the New Year's Eve topic were the majority, along with minor topics that could also be found in the news such as the new law for smoking in public places. Although results for one day are promising, topics roughly can be taken from here, since the mix of hot terms for one day appears to be a bit chaotic. With the need to select a large period of time, we decided to go one step further and detect topical trends during one month.

4.4 Detecting Trends over Time

After preparing the collection for trend detection, we can put in practice the methodology described in Section 4.1 for detecting important terms in a selected period of time. Our objective here is to be able to detect for a given month the main topics that were discussed in the portuguese Blogspace. If the methodology proves to be successful, we can then automatize the model so that trends can be delivered for each month of our time span.

4.4.1 Extracting Important Terms

Following the Frequency Segments algorithm, we calculate the score for each one of the 105,231 terms covered by our detector. Table 4.4 presents the first results obtained using this algorithm for December 2007. After retrieving this results, we search for our conclusions trying to relate terms that are described by our mechanism as important terms with events that were covered by the media in the same time span.

federation	google.com	bhutto	russian
benazir	bucetinha	venezuelanos	kadhafi
musharraf	mugabe	líbia	fjv
louletano	niemeyer	sudão	lip
assassinada	maxi	offers	426
navidad	apito	brincarem	accionistas
unicef	venezuelana	assassinatos	bali
treva	luisão	imaculada	cody
21h30	reformador	petitiononline.com	articulação
armazém	cmvm	esfrega	prefeitura

Tabela 4.4: Relevant terms using frequency segments algorithm ($\alpha = 0.33$)

Although results presented in Table 4.4 appeared to be reasonable, it is visible that by applying a methodology tested in a small dataset (40,000 entries) to a large corpus with over 3,000,000 posts, two main problems were introduced:

- Since we use the daily DF to calculate the score of a term, without cross-checking the source of these values, terms coming from splogs can be easily introduced when you are considering a larger dataset. This means that terms used by blogs to increase their page rank for some search queries can be erroneously detected by our mechanism as important topics for one month. Terms that should be considered as spam for December 2007 appear in light gray in Table 4.4; we consider as spam all those terms that either repeatedly came from the same blog or from posts with the same title and content in the exact same day.

- Also, as we were expecting, terms that occurred everyday during one month were not considered by the Frequency Segments algorithm. Although results seemed to be representing well topics that were important for December 2007, we think that changes in the algorithm can be made so that terms that appear everyday can also be considered relevant for a defined period, if that is the case; for this month, it would be usual for terms related to the Christmas season to show up among the most relevant terms.

Figure 4.9 shows the frequency of occurrences for the top 3 terms detected by the Frequency Segment algorithm for December 2007; here, it is possible to witness how the frequencies for the three terms has increased dramatically during the first half of December, thus being correctly considered as a trend if we just took into consideration the DF values.

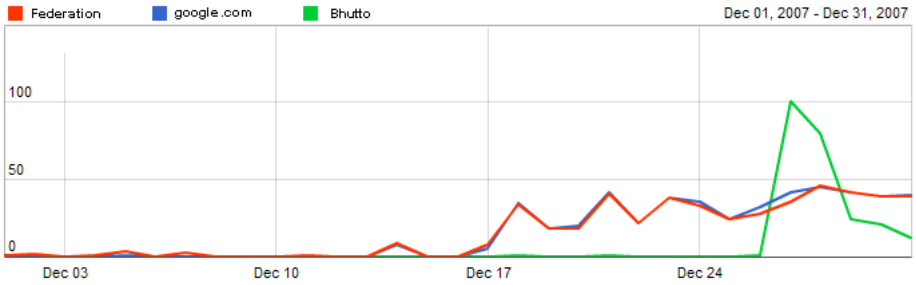

Figura 4.9: Trend for “federation”, “google.com” and “bhutto” in December 2007

4.4.2 Spam Removal

As we already mention in Chapter 2, splogs have been frequently studied in an attempt to cull them out of the Blogspace. Although spam removal is not in the domain of this work, neither having been referred by Oka et al. in their paper, the presence of results originated by spam blogs was big enough to deserve our attention, since this was clearly interfering with clear results for trends detection.

After inspecting results for each of the 40 terms retrieved for December 2007, we found out that spam results that were wrongly detected by our system were always produced either by the same blog (posting a substantial amount of posts with the same word over and over) or by the same post (same title and content, although coming from different blogs). This can be seen in Figure 4.10, where the top results for the term “federation” during December 2007 are shown using Luke[4], a toolbox used to inspect a Lucene index. The bottom half of the picture shows how top results came in the same day from the same blog.

[4]http://www.getopt.org/luke/

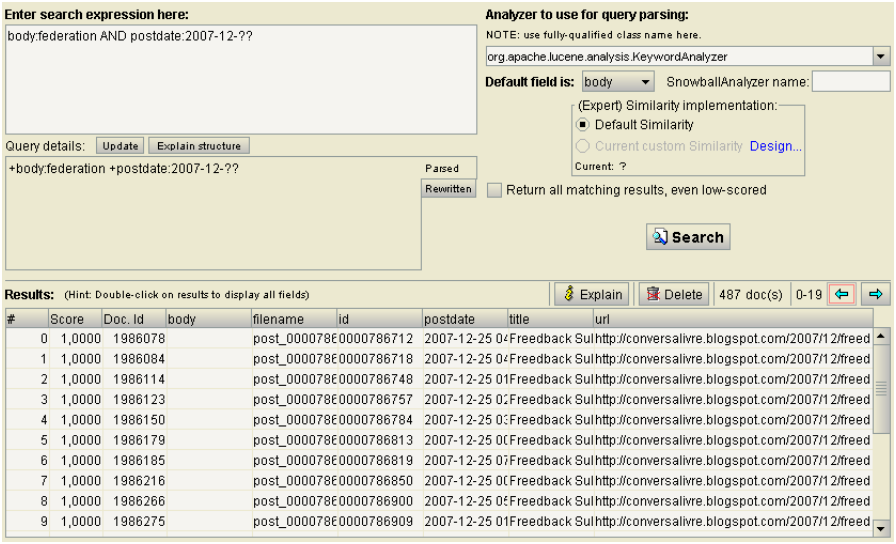

Figura 4.10: Results for “federation” for December 2007

To solve this problem, we decided to filter all the results with a simple algorithm that only keeps those terms in which, relatively to the day of the given month where their DF is maximum, at least half of the posts are both from a different blog and present different titles. This simple mechanism permitted to retrieve a final list of important terms, presented in Table 4.5.

bhutto	benazir	venezuelanos	kadhafi
musharraf	mugabe	líbia	louletano
niemeyer	sudão	assassinada	maxi
426	navidad	apito	accionistas
unicef	venezuelana	assassinatos	bali
treva	luisão	imaculada	cody
21h30	reformador	petitiononline.com	articulação
armazém	cmvm	prefeitura	estabelece
renan	isento	compotas	eurostat
bastonário	2-3	0-2	cadilhe

Tabela 4.5: Relevant terms for December 2007 after spam remotion

4.4.3 Terms Occurring Everyday

As already explained in Section 4.1, the Frequency Segments algorithm uses zeros in the sequence of occurrences to establish the boundaries of segments that will later be compared and ranked. This means that if a term only has one segment, by occurring in sequent days without any intervals, the methodology proposed by Oka et al. will not consider this term for ranking. Although this was a choice done by the researchers who developed the algorithm, in our case ignoring these terms would have been ignoring some of the more

important terms that occurred in the selected period of time.

Instead of developing a new methodology, we want to adapt the Frequency Segments algorithm with one minor change that would enable the existing methodology to capture terms that only occur in sequent days. In their work, Oka et al. decide to form a new segment of occurrences each time a term has $DF = 0$, and although this ensures that segments will describe the period in which term T was active in the Blogspace, the mathematical formulation of the problem is still valid if we force the creation of segments, dividing the period of time in similar fractions, each one representing a fragment. Every segment would have the same length, but with autonomous values for segment sum and deviation.

In our algorithm, we employed a method that forces a segment every 6 days, given that a month has around 30 days and being this division the one that would assure larger values of segment sum and deviation sum to be detected. The idea was that although sometimes a segment could not represent the entire period in which a term was possibly important, it would certainly detect a big fraction of this period.

In a first step, we employed the two algorithms together. Our method is an extension of the traditional one that is only used each time a term does not have more than one segment for the defined period. Table 4.6 shows the results achieved with this modification.

natal	feliz	cimeira	tratado
festas	áfrica	novembro	bhutto
venezuela	chávez	benazir	santo
desejar	venezuelanos	africanos	kadhafi
musharraf	mugabe	líbia	louletano
bcp	deseja	ue	christmas
niemeyer	sudão	assinatura	hugo
merry	derrota	quadra	próspero
assassinada	maxi	taxa	426
navidad	apito	accionistas	unicef

Tabela 4.6: Relevant terms for December 2007 after introducing values with zeros

Comparing to the previous results, terms included by our change in the algorithm — as “cimeira” (summit) and “natal” (Christmas) — are visibly more common terms then the ones previously presented. In our ranking of 40 terms for December 2007, 27 were new; when inspecting the news archive for the same time period, results seemed to represent reality. Figure 4.11 shows the frequency of occurrences for some of the terms that ranked higher after the inclusion of results with one segment only: “natal”, “feliz”, “cimeira” and “bhutto”.

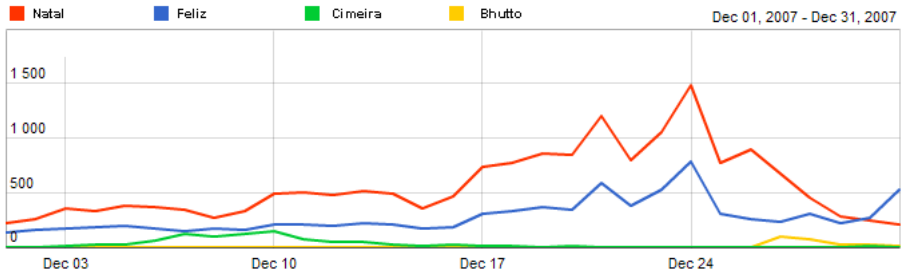

Figura 4.11: Trend for “natal”, “feliz”, “cimeira” and “bhutto” in December 2007

But while the two methods for creating segments worked well together, we still have to know if our change in the detection of segments affects results when working alone. Our next step is to create the ranking for December 2007 once again, only this time every term would be scored with values for the same segments, representing a division of a month every 6 days, instead of recurring to zeros in the frequency of occurrences. Results are presented in Table 4.7.

natal	feliz	cimeira	benazir
tratado	festas	bhutto	áfrica
novembro	venezuela	chávez	santo
desejar	paquistão	africanos	mugabe
bcp	deseja	ue	christmas
assinatura	hugo	chavez	merry
derrota	cgd	quadra	próspero
niemeyer	venezuelano	taxa	reformador
ditadores	venezuelanos	musharraf	darfur
presépio	cpmf	referendo	kadhafi

Tabela 4.7: Relevant terms recurring only to weekly segments

After employing only our method for segmentation, we needed to understand what were the exact changes that our algorithm introduced. Tables 4.8 and 4.9 refer to the terms that were included and excluded respectively; values on the table represent values for our segmentation algorithm, while values in parenthesis represent values for the old algorithm. After inspecting these results, we can get to some simple conclusions:

- Segmentation by Zeros: the more segments, the bigger is the segment deviation if a value is important. If the term has fewer segments, the value will be discarded. This happened with the terms “paquistão” and “ditadores”, that were discarded by the first algorithm because each one only had one or two segments (although important ones)

- Segmentation by Month Fractions: values for segment deviation will be smaller, since less segments will appear, thus not giving so much importance to words that

don't often appear in the corpus; these terms are only shown if they have a high segment sum. That way, terms like "benazir" and "bhutto" were kept, although they had 6 segments each, but the co-occurring term "assassinada" was not.

term	average segment	segment sum	segment deviation	maximum	zeros
líbia	12.6 (9.27)	76 (102)	0.23 (0.56)	23	6
louletano	13.5 (7.4)	81 (111)	0.28 (0.51)	34	5
sudão	11.1 (6.4)	67 (128)	0.19 (0.41)	19	4
assassinada	11.6 (12.5)	70 (75)	0.40 (0.52)	19	6
maxi	11.3 (7.0)	68 (149)	0.12 (0.29)	35	3
426	5.6 (6.8)	34 (48)	0.21 (0.66)	19	8
navidad	7.6 (4.7)	46 (81)	0.21 (0.49)	17	5
apito	11.3 (9.8)	68 (237)	-0.14 (0.13)	20	2
accionistas	13.8 (10.9)	83 (131)	0.19 (0.32)	31	4
unicef	5.8 (5.1)	35 (62)	0.07 (0.57)	13	7

Tabela 4.8: Terms Excluded from the Ranking

term	average segment	segment sum	segment deviation	maximum	zeros
paquistão	30 (24.1)	180 (193)	0.41 (-0.15)	77	1
cgd	22.8 (12.5)	137 (201)	0.34 (-0.06)	34	1
venezuelano	22.1 (13.3)	133 (173)	0.33 (0.07)	49	2
reformador	17.6 (9.2)	106 (139)	0.39 (0.28)	41	3
ditadores	22.6 (10.0)	136 (211)	0.31 (-0.05)	31	1
darfur	23.6 (23.6)	142 (142)	0.27 (0.27)	44	0
presépio	50 (50)	300 (300)	0.01 (0.01)	64	0
cpmf	30.3 (30.3)	182 (182)	0.18 (0.18)	73	0
referendo	45.6 (45.6)	274 (274)	0.02 (0.02)	102	0

Tabela 4.9: Terms Included in the Ranking

We think that the creation of frequency segments using weeks as well defined segments, instead of recurring to zeros in the frequency of occurrences, gives more importance to those terms that were more talked about during the defined period, rather than the previous method of segmentation that gives a higher ranking to terms that appear fewer times, with many different sequent segments. To help understand the terms chosen over others, we compare in Figures 4.12 and 4.13 the frequency of occurrences of two terms that were included after the algorithm is changed ("paquistão" and "referendo") and two terms that were excluded from the top 40 ("assassinada" and 'líbia").

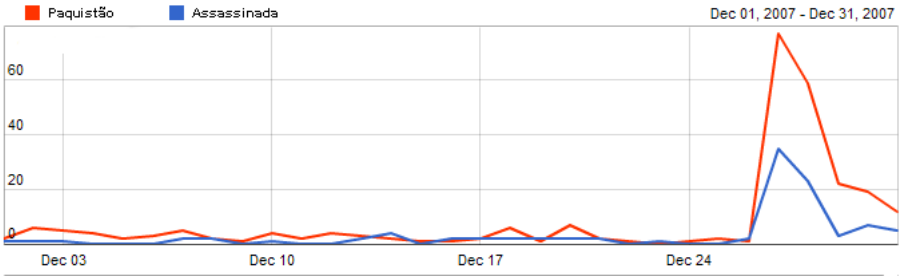

Figura 4.12: Trend for “paquistão” and “assassinada” in December 2007

In Figure 4.12 the two terms compared are co-occurrent in the corpus, since they both are related to the assassination of former Pakistan prime minister Benazir Bhutto. Here, the difference between the two methodologies is very visible, since with the segmentation with zeros the term “paquistão” would not have been chosen, given it already had 2 segments, thus obtaining a lower result of segment deviation when compared with others like ‘líbia”, with 7 segments.

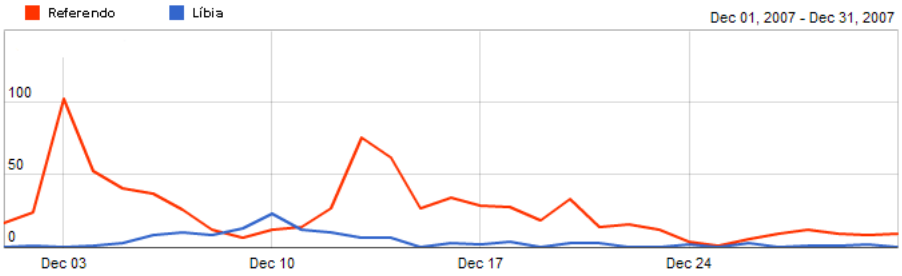

Figura 4.13: Trend for “referendo” and “líbia” in December 2007

Figure 4.13 depicts “referendo”, the lowest ranked included term and “líbia”, the highest ranked excluded term, graphically concluding that although “líbia” was also part of a main topic in the first half of December, “referendo” was definitely a more important trend for December 2007, thus making sense that this term appears on the results, on the top of all the others that were excluded.

We do not want to conclude that our method for selecting segments is better than the one originally employed in the Frequency Segments algorithm, since there is no doubt that terms retrieved with it were part of hot topics during the month. However, we think that while searching for the most important trends in one month rather than the most important topics that popped for two or three days in the whole month, using equal segments for every term achieves better results.

Therefore, we think it’s possible we are facing a possible two-tier mechanism that could be used to first retrieve the more frequently written terms, and then relate them with

the ones that appear with the original segmentation, the first being more topic-like and the others having a much more descriptive nature.

4.4.4 Term Clustering

Like Oka et al., we grouped similar terms together, therefore forming more descriptive topics that would hopefully be representative of the month. We want to group together co-occurrent terms like the ones depicted by Figure 4.14.

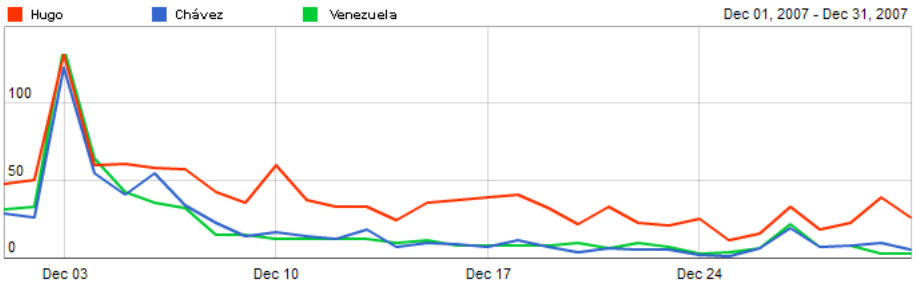

Figura 4.14: Trend for “hugo”, “chávez” and “venezuela” in December 2007

Our method was slightly different from Oka et. al, since instead of first selecting all of the terms that co-occurred in the same post and only then rank them by their correlation of frequencies, we decided that it would be appropriate to form a score describing how well words are connected, through their score of correlation (Equation 4.4 in Section 4.1) and their diversity in the corpus. For each two terms A and B, our score is calculated recurring to Equation 4.6; $N_{A,B}$ represents the number of posts in which A and B appear together, while N_A represents the total number of entries with term A. The first part of the equation represents the correlation between the vectors of occurrence of terms A and B in the selected period. However, two terms having highly correlated vectors of occurrence may not represent the same topic. Therefore, we need the second part of the equation to ensure that even if two terms are highly correlated they can only have high similarity score if they often occur together in the same documents.

$$similarity(A,B) = \frac{< A - mean(A), B - mean(B) >}{\|A - mean(A)\| \times \|B - mean(B)\|} \times \frac{N_{A,B}}{N_A} \tag{4.6}$$

After selecting the 20 top terms for December 2007, we calculated the similarity level for each one of them with all the others. If an empirically determined threshold of $similarity(A,B) = 1.5$ was passed, we would join term B in the cluster for term A. We calculated a similarity value for all the combinations of the last 20 terms ranked as important terms for December 2007, clustering every term B that has a threshold of $similarity(A,B) = 1.5$ above which similar terms were kept.

ranking	term	related terms
1	natal	feliz
2	feliz	natal
3	cimeira	ue, mugabe, africanos, áfrica
4	benazir	bhutto, paquistão
5	tratado	ue
6	festas	feliz, natal
7	bhutto	paquistão, benazir
8	áfrica	cimeira, ue, africanos
9	novembro	
10	venezuela	chávez
11	chávez	venezuela
12	santo	natal, feliz
13	desejar	feliz, natal
14	paquistão	bhutto, benazir
15	africanos	mugabe, áfrica, cimeira, ue
16	mugabe	áfrica, africanos, ue, cimeira
17	bcp	
18	deseja	feliz, natal
19	ue	áfrica, tratado, cimeira
20	christmas	natal

Tabela 4.10: Top 20 terms for December 2007 and related terms

After all similar terms are related, each of the clusters of terms was recursively grouped with other that had more than half of the terms similar to one another. Results are shown on Table 4.11.

ranking	topic
1	natal feliz desejar deseja
2	natal christmas
3	natal feliz festas santo
4	cimeira tratado áfrica africanos mugabe ue
5	benazir bhutto paquistão
6	venezuela chávez

Tabela 4.11: Top 20 terms clustered

We think that the values obtained for this month as well as the others that were selected by our algorithm show cluster of terms, or topics, that can assure they are representative, and in a certain way descriptive, of what happens during the month in question. Obviously, when the number of terms that are used as an input to the clustering mechanism increases, one can have more insight on the topics talked about in the selected period.

ranking	topic
1	natal feliz festas santo deseja próspero
2	natal feliz desejar quadra
3	natal feliz christmas merry
4	natal feliz presépio
5	cimeira tratado áfrica ue
6	cimeira áfrica africanos mugabe ue ditadores darfur kadhafi
7	benazir bhutto paquistão musharraf
8	tratado venezuela chávez hugo referendo
9	tratado ue assinatura reformador referendo
10	venezuela chávez hugo chavez derrota venezuelano venezuelanos referendo
11	bcp cgd

Tabela 4.12: Top 40 terms clustered

Table 4.12 shows the topics that were selected as the most important for December 2007; with a simple inspection using Google News Archive[5], we are able to manually trace a connection between these topics and the ones presented in the media by December 2007:

- Topics 1, 2, 3, 4 and 5 refer to the Christmas season, encompassing other words associated with this holiday;
- Topic 5 and 6 refer to the EU-Africa Summit, held for the first time in Portugal during the month of December;
- Topic 7 refers to the assassination of Benazir Bhutto, former Pakistan Prime Minister;
- Topic 8 and 10 refer to the defeat of Hugo Chávez in the referendum to change the constitution laws in Venezuela;
- Topic 9 refers to the Treaty of Lisbon, signed in Portugal;
- Topic 10 refers to the replacement of leadership in the portuguese bank Banco Comercial Português by the old Caixa Geral de Depósitos' CEO.

4.5 Summary and Conclusions

In this section we show how it is possible to adapt existing methodologies to a large portuguese corpus. Although the Frequency Segments algorithm showed to be capable of retrieving important topics for our dataset, results were poor considering the large amount

[5]http://news.google.com/archivesearch

of spam introduced and the lack of terms that are commonly written everyday.

With a change in the paradigm of selecting the segments with which the Frequency Segments algorithm works, creating segments by predefined cuts in the selected period instead of segmenting results by their zeros, together with the introduction of a simple filter for spam words, we were able to produce results that match an informal evaluation of what has been present in blogs during the past year in Portugal.

In the end, we are faced with two mechanisms that can retrieve the most important terms for both a month and a day, thus giving us the possibility to create a well defined archive of what was talked in the portuguese Blogspace in recent years; Appendix B presents a compilation of 2007's more important terms and topics retrieved by our system.

Capítulo 5

Conclusions

With this work, we try to prove that it is possible to adapt an existing methodology employed with success in previous Blog Trend Detection research so that it performs appropriately on a complex portuguese blog corpus. Beginning with a large overview about what exactly is a blog and its potential as a research field, we follow with a characterization of the dataset made available by the portuguese company SAPO. We then use this collection to automatically retrieve hot terms and important topics that were discussed in the portuguese Blogspace over the past year. In this final chapter, we do a short summary of our contributions, and define directions for future work.

5.1 Summary of Contributions

Even though our main objective was to study detection methods for blog trends, our work is divided into two parts: the characterization of a portuguese blogs dataset and the study of automatic detection of trends in the portuguese Blogspace. We start to characterize the dataset given by SAPO because this was the first opportunity to analyze a portuguese blog collection and present it as scholarly work. Also, this characterization gave us a better insight about the contents of the collection we would then use to infer topical discussions.

Characterization of Portuguese Blogs Dataset With 49,940 blogs and 2,933,735 posts spanning 5 years between January 2003 and December 2007, the collection presented blog entries from two main blog providers: SAPO Blogs (52%) and Blogpsot (47%). Assured to be composed by posts mostly written in portuguese from Portugal, this collection shows a stable evolution of the number of blogs, with slowdowns on April 2006 and February 2007, possibly related with the SAPO crawler. Regarding posts, a sudden growth of posting activity was found in April 2006. We think that the growths and slowdowns on blog creation and posting activity of this collection cannot be directly related to the ones experienced by the entire portuguese Blogspace, since these values can be mislead by

the work of the SAPO crawler. However, due to its considerable size, discussions, word usage and link creation in this dataset are probably according to what was published by Portuguese bloggers between 2003 and 2007.

Automatic Detection of Trends in the Portuguese Blogspace Having many methodologies for trend detection in the Blogspace from which to choose, the Frequency Segments algorithm appeared to be the one with the most simple application that still achieved good results. Recurring to a simple count of the frequency of occurrences for each term, this algorithm proved to be effective even when applied to our large collection. But this was not free of problems: spam terms occurred in every rank of important terms and terms that appear everyday in the collection were not considered. Recurring to a simple spam filter and by changing the method to calculate segments for each term, we came up with a variation of the Frequency Segments algorithm that achieved better results when looking for well recognizable topics in a selected period of time. In the end, we are able to present two mechanisms that can retrieve the most important terms for both a month and a day, giving a better insight of what happened in the portuguese Blogspace during the last year.

5.2 Future Work

We present some guidelines for future work that can be developed on the top of our research. This is work that either we did not have time to presume or that was considered as being outside the scope of this dissertation.

Evaluation of Results Although the results achieved by our trend detection mechanism seemed to match an informal evaluation of what has been present in blogs during the past year in Portugal, our work still lacks a precise evaluation that could determine the main differences between the usage of the Frequency Segments algorithm as proposed by Oka et al. and the usage of the same algorithm with the changes introduced by us. It would be challenging to simulate values for the occurrences of terms, determine what results would be expected and see how the two detection mechanisms would work.

Entity and Phrase Finding Instead of detecting hot terms and important topics only based on the frequency of n-grams of size 1, it would be interesting to create the same file set considering entities and phrases, that could than be ranked as important terms by the same methodology. This would deliver more descriptive topics, that might be better understood by users.

Automatically Linking News Articles with Blog Entries From where we stand, it would not be difficult to link each one of the topics retrieved by our clustering algorithm to

results presented in the mass media. One of our basic evaluation of results was to query the Google News Archive[1] with each one of our cluster of terms and the period selected. Automating this kind of actions would not be a hard task.

Detection of Communities and Authorities By correlating plots of different trends, we could try to create clusters of news divided by theme: politics, sports, science, etc. Another possibility would be to create user communities that were specific to a set of news articles and try to distinguish authority blogs that usually give the news first than others.

Visualization Tool All plots presented in the document were created recurring to the AmCharts set of Flash Charts[2]. Although these charts worked as expected, the creation of our own visualization tool would allow us to have a better control of how results were presented and the kind of interaction users could have with them. Rather than just presenting the chart of a trend for a topic, we could specify what kind of discussion was related to each of the bursts in a plot.

Automation of Real Word Mechanism Finally, it would be interesting to compile all this work together with a crawler developed by us that would present daily results of today's hot terms and important topics. After this dissertation, we feel that it is not a hard task to proceed to an automatic mechanism that would inspect the portuguese Blogspace on a daily basis.

[1] http://news.google.com/archivesearch
[2] http://www.amcharts.com/

Bibliografia

[1] WorldWideWebSize.com, The size of the World Wide Web . National Vulnerability Database, June 2008. [online] http://www.worldwidewebsize.com/.

[2] Brian Ulicny, Ken Baclawski, and Amy Magnus. New metrics for blog mining. In *Proceedings of SPIE Defense & Security Symposium*, 2007.

[3] 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.

[4] G. Mishne. *Applied Text Analytics for Blogs*. PhD thesis, University of Amsterdam, 2007.

[5] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. the bursty evolution of blogspace. In *Proceedings of WWW 2005*. Springer, 2005.

[6] D. Gruhl, R. Guha, Liben D. Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web*, New York, NY, USA, 2004. ACM.

[7] Mukul Joshi and Nikhil Belsare. Blogharvest: Blog mining and search framework. In *Proc. of the Int'l Conf. on Management of Data COMAD*, 2006.

[8] Fujimura. The eigenrumor algorithm for ranking blogs. In *WWW2005 Workshop on the Weblogging Ecosystem*, 2005.

[9] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[10] Wu. Important weblog identification and hot story summarization. In *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.

[11] Adar. Implicit structure and the dynamics of blogspace. In *Workshop on the Weblogging Ecosystem*, 2004.

[12] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[13] A. Kritikopoulos, M. Sideri, and I. Varlamis. Blogrank: ranking weblogs based on connectivity and similarity features. In *Proceedings of the 2nd international Workshop on Advanced Architectures and Algorithms for Internet Delivery and Applications*, 2006.

[14] S. Nakajima, J. Tatemura, Y. Hino, Y. Hara, and K. Tanaka. Discovering important bloggers based on analyzing blog threads. In *Proceedings of WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem*, 2005.

[15] Paula Chesley, Bruce Vincent, Li Xu, and Rohini Srihari. Using verbs and adjectives to automatically classify blog sentiment. In *AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.

[16] L. W. Ku, Y. T. Liang, and H. H. Chen. Opinion extraction, summarization and tracking in news and blog corpora. In *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.

[17] T. Mullen and R. Malouf. A preliminary investigation into sentiment analysis of informal political discourse. In *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.

[18] G. Mishne. Experiments with mood classification in blog posts. In *1st Workshop on Stylistic Analysis Of Text For Information Access*, 2005.

[19] G. Mishne and N. Glance. Predicting movie sales from blogger sentiment. In *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.

[20] G. Mishne. Using blog properties to improve retrieval. In *ICWSM'2007*, Boulder, Colorado, USA, 2007.

[21] Gilad Mishne. Multiple ranking strategies for opinion retrieval in blogs. In *TREC*, Gaithersburg, Maryland USA, 2006.

[22] Akshay Java, Pranam Kolari, Tim Finin, and Tim Oates. Modeling the spread of influence on the blogosphere. In *WWW 2006 Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2006.

[23] Pranam Kolari, Tim Finin, and Anupam Joshi. Svms for the blogosphere: Blog identification and splog detection. In *AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.

[24] Kazuyuki Narisawa, Yasuhiro Yamada, Daisuke Ikeda, and Masayuki Takeda. Detecting blog spams using the vocabulary size of all substrings in their copies. In *WWW 2006 Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2006.

[25] Matthew Hurst. 24 hours in the blogosphere. In *AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.

[26] Seungyeop Han, Yong Y. Ahn, Sue Moon, and Hawoong Jeong. Collaborative blog spam filtering using adaptive percolation search. In *InWWW 2006 Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2006.

[27] Natalie Glance, Matthew Hurst, Kamal Nigam, Matthew Siegler, Robert Stockton, and Takashi Tomokiyo. Deriving marketing intelligence from online discussion. In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 2005.

[28] Richard M. Tong and Mark Snuffin. Weblogs as market indicators: Tracking reactions to issues and events. In *AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.

[29] Daniel Gruhl, Ramanathan Guha, Ravi Kumar, Jasmine Novak, and Andrew Tomkins. The predictive power of online chatter. In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 2005.

[30] Andreas Aschenbrenner and Silvia Miksch. Blog mining in a corporate environment. Technical report, SAT, 2005.

[31] Matthew Hurst. Temporal text mining. In *AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.

[32] M. Oka, H. Abe, and K. Kato. Extracting topics from weblogs through frequency segments. In *WWW2006*, Edinburgh, UK., May 2006.

[33] Tomohiro Fukuhara, Toshihiro Murayama, and Toyoaki Nishida. Analyzing concerns of people using weblog articles and real world temporal data. In *WWW 2005 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2005.

[34] Levon Lloyd, Prachi Kaulgud, and Steven Skiena. News vs. blogs: Who gets the scoop? In *AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.

[35] Natalie S. Glance, Matthew Hurst, and Takashi Tomokiyo. Blogpulse: Automated trend discovery for weblogs. In *WWW 2004 Workshop on the Weblogging Ecosystem*, New York, NY USA, 2004. ACM.

[36] Chun-Yuan Teng and Hsin-Hsi Chen. Detection of bloggers' interests: Using textual, temporal, and interactive features. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 366–369, Washington, DC, USA, 2006. IEEE Computer Society.

[37] Arun Qamra, Belle Tseng, and Edward Y. Chang. Mining blog stories using community-based and temporal clustering. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 58–67, New York, NY, USA, 2006. ACM Press.

[38] Yun Chi, Belle L. Tseng, and Junichi Tatemura. Eigen-trend: Trend analysis in the blogosphere based on singular value decompositions. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, Arlington, Virginia, USA, 2006. ACM.

[39] Ding Zhou, Xiang Ji, Hongyuan Zha, and Lee C. Giles. Topic evolution and social interactions: How authors effect research. In *CIKM '06: Proceedings of the fifteenth international conference on Information and knowledge management*, 2006.

[40] Yi Wu and Belle L. Tseng. Important weblog identification and hot story summarization. In *AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.

[41] M. Gamon, S. Basu, D. Belenko, D. Fisher, and M. Hurst. Blews: Using blogs to provide context for news articles. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*. Association for the Advancement of Artificial Intelligence, 2008.

[42] R. Nallapati and W. Cohen. Link-plsa-lda: A new unsupervised model for topics and influence of blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*. Association for the Advancement of Artificial Intelligence, 2008.

[43] José Castelo Branco. Aplicação do h-index em blogues. Master's thesis, Faculdade de Engenharia da Universidade do Porto, July 2008.

[44] Craig Macdonald and Iadh Ounis. The trec blogs06 collection : Creating and analysing a blog test collection. Technical report, Department of Computing Science, University of Glasgow, 2006.

[45] Vahed Qazvinian, Abtin Rassoliau, and Mohammad Shafiei. A large-scale study on persian weblogs. In *Proc. of Workshop on Text-Mining and Link-Analysis*, 2007.

[46] P. Kolari, A. Java, and T. Finin. Characterizing the splogosphere. In *Proceedings of the 3rd Annual Workshop on the Weblogging Ecosystem*, 2006.

Apêndice A

Relevant Blogs and Entries

While characterizing our dataset, we retrieved some listings that may be interesting under a social context. Although results should not be taken as a strict description of what reality is, we can say that they can be a good indicator of what were, for instance, the most important blogs and posts in the portuguese blogspace between 2003 and 2007.

A.1 Blogs by Number of Posts

ranking	blog	nr posts
1	opiolhodasolum.blogspot.com	8378
2	cartunistasolda.blogspot.com	8225
3	www.planetgeek.org	6946
4	angelorigon.blogspot.com	6302
5	corta-fitas.blogspot.com	5286
6	amizadeverdadeira-amizade.blogspot.com	4667
7	ablasfemia.blogspot.com	4557
8	carvalhadas-on-line.blogspot.com	4482
9	domelhor.net	4341
10	braganza-mothers.blogspot.com	4073
11	atonito.blogspot.com	4036
12	mafiadacova.blogspot.com	3953
13	ultraperiferias.blogspot.com	3859
14	conversalivre.blogspot.com	3635
15	santosdacasa.blogspot.com	3578
16	attu.blogspot.com	3534
17	macroscopio.blogspot.com	3458
18	pavablog.blogspot.com	3437
19	alhosvedrosaopoder.blogspot.com	3436
20	atrompa.blogspot.com	3302
21	asterisco.paradigma.pt	3198
22	sorumbatico.blogspot.com	3064
23	lestedeangola.weblog.com.pt	3043
24	antologiadoesquecimento.blogspot.com	2984
25	coisasbobas.blogspot.com	2959
26	diasquevoam.blogspot.com	2934
27	lagrimapsicodelica.blogspot.com	2842
28	culinaria.weblog.com.pt	2800
29	ouro-sobre-estrelas.blogspot.com	2777
30	www.domelhor.net	2741
31	comunicaradireito.esta.weblog.com.pt	2739
32	siter.blogs.sapo.pt	2654
33	cortenaaldeia.blogspot.com	2595
34	timor-online.blogspot.com	2581
35	preca.blogspot.com	2542
36	kropotkine.blogspot.com	2521
37	oviseu.blogspot.com	2471
38	geracao-rasca.blogspot.com	2425
39	portugaldospequeninos.blogspot.com	2408
40	corporacoes.blogspot.com	2392
41	abrupto.blogspot.com	2379
42	kostadealhabaite.blogspot.com	2334
43	bloguitica.blogspot.com	2334
44	alcatruz.blogspot.com	2294
45	jansenista.blogspot.com	2281
46	riotorelse.blogspot.com	2277
47	cidadanialx.blogspot.com	2259
48	incontinentesverbais.blogspot.com	2228
49	guitarradecoimbra.blogspot.com	2226
50	pardalitosdochoupal.blogspot.com	2212

Tabela A.1: Ranking of Blogs by Number of Posts

A.2 Blogs by Incoming Links

ranking	blog	inlinks	self inlinks	dif
1	ablasfemia.blogspot.com	4215	128	4087
2	abrupto.blogspot.com	2994	136	2858
3	causa-nossa.blogspot.com	1867	99	1768
4	daliteratura.blogspot.com	1711	216	1495
5	origemdasespecies.blogspot.com	1745	272	1473
6	elvirabistrot.blogspot.com	1678	270	1408
7	portugaldospequeninos.blogspot.com	1586	238	1348
8	corta-fitas.blogspot.com	1558	222	1336
9	bloguitica.blogspot.com	1459	212	1247
10	gloriafacil.blogspot.com	1161	30	1131
11	bomba-inteligente.blogspot.com	1129	144	985
12	estadocivil.blogspot.com	976	3	973
13	doportugalprofundo.blogspot.com	819	0	819
14	oinsurgente.blogspot.com	936	124	812
15	tugir.blogspot.com	1003	206	797
16	combustoes.blogspot.com	748	23	725
17	bichos-carpinteiros.blogspot.com	700	5	695
18	wehavekaosinthegarden.blogspot.com	687	25	662
19	aartedafuga.blogspot.com	1088	430	658
20	grandelojadoqueijolimiano.blogspot.com	734	79	655
21	blogueforanada.blogspot.com	646	21	625
22	incursoes.blogspot.com	650	43	607
23	odoloeventual.blogspot.com	1048	445	603
24	hojehaconquilhas.blogspot.com	628	47	581
25	misspearls.blogspot.com	664	89	575
26	acausafoimodificada.blogspot.com	555	2	553
27	geracao-rasca.blogspot.com	812	291	521
28	luiscarmelo.blogspot.com	516	44	472
29	macroscopio.blogspot.com	563	98	465
30	portugalcontemporaneo.blogspot.com	479	20	459
31	quartarepublica.blogspot.com	576	118	458
32	avenida-dos-aliados-porto.blogspot.com	489	32	457
33	antologiadoesquecimento.blogspot.com	36493	36050	443
34	ocanhoto.blogspot.com	483	40	443
35	filhodo25deabril.blogspot.com	496	73	423
36	anaturezadomal.blogspot.com	411	0	411
37	jansenista.blogspot.com	398	2	396
38	blogotinha.blogspot.com	389	7	382
39	eroticidades.blogspot.com	384	6	378
40	edeuscriouamulher.blogspot.com	415	48	367
41	oacidental.blogspot.com	375	12	363
42	tempoquepassa.blogspot.com	375	12	363
43	esplanar.blogspot.com	359	2	357
44	avenidacentral.blogspot.com	414	77	337
45	quaseemportugues.blogspot.com	333	0	333
46	dias-com-arvores.blogspot.com	373	40	333
47	encarnados.blogspot.com	612	289	323
48	divasecontrabaixos.blogspot.com	531	208	323
49	frenchkissin.blogspot.com	353	31	322
50	margensdeerro.blogspot.com	413	105	308

Tabela A.2: Ranking of Blogs by Incoming Links

A.3 Blogs by Self Incoming Links

ranking	blog	self inlinks	inlinks
1	antologiadoesquecimento.blogspot.com	36050	36493
2	antoniopovinho.blogspot.com	2049	2276
3	lagrimapsicodelica.blogspot.com	1901	1932
4	seehere.blogspot.com	1767	1780
5	lampadamagica.blogspot.com	1223	1254
6	intergalacticrobot.blogspot.com	1178	1198
7	veraoverdeorg.blogspot.com	1101	1128
8	ruyluisgomes.blogspot.com	999	1101
9	kouzaselouzas.blogspot.com	840	845
10	vouatuacasa.blogspot.com	833	910
11	bretemas.blogspot.com	779	933
12	atrompa.blogspot.com	763	843
13	gonn1000.blogspot.com	706	749
14	jorgenunopintodacosta.blogspot.com	693	722
15	kraakfm.blogspot.com	617	719
16	criticademusica.blogspot.com	604	612
17	minisaia.blogspot.com	601	689
18	amygdalagf.blogspot.com	587	587
19	devaneiosdesintericos.blogspot.com	574	776
20	lerbd.blogspot.com	563	633
21	praiadosmoinhos.blogspot.com	537	545
22	camaradapatrao.blogspot.com	533	537
23	assumidamente.blogspot.com	486	546
24	amornatural.blogspot.com	480	481
25	odesproposito.blogspot.com	455	527
26	desnorte.blogspot.com	451	494
27	a-sul.blogspot.com	450	521
28	juliosevero.blogspot.com	447	521
29	odoloeventual.blogspot.com	445	1048
30	lagrimadedor.blogspot.com	439	468
31	irrealtv.blogspot.com	439	501
32	aartedafuga.blogspot.com	430	1088
33	ismspt.blogspot.com	427	427
34	tiago_ribeiro.blogspot.com	410	410
35	indeterminacy.blogspot.com	406	408
36	corporacoes.blogspot.com	400	708
37	mitos-climaticos.blogspot.com	390	459
38	umamallapelomundo.blogspot.com	385	462
39	dofundodomar.blogspot.com	383	383
40	povodebaha.blogspot.com	378	451
41	nothingandall.blogspot.com	371	422
42	writeinwater.blogspot.com	356	359
43	gavetadenuvens.blogspot.com	349	351
44	biscoitos-terceira.blogspot.com	348	350
45	santamargarida.blogspot.com	346	363
46	liberallibertariolibertino.blogspot.com	336	524
47	lorenzetti.blogspot.com	332	349
48	parolesimages.blogspot.com	331	341
49	pedraformosa.blogspot.com	315	336
50	periplus.blogspot.com	306	309

Tabela A.3: Ranking of Blogs by Self Incoming Links

A.4 Posts by Incoming Links

ranking	post	inlinks
1	http://gloriafacil.blogspot.com/2006/09/arrojadamente-com-arrojo-etc.html	124
2	http://geracao-rasca.blogspot.com/2006/11/regulamento.html	87
3	http://doportugalprofundo.blogspot.com/2007/06/arguido-por-causa-do-dossier-scrates.html	54
4	http://encarnados.blogspot.com/2006/07/crnicas-encarnadas-200607.html	95
5	http://combustoes.blogspot.com/2007/04/fassismo-feixismo-e-outras-prepotncias.html	39
6	http://socioeleicoes.blogspot.com/2006/05/valorar-absteno-como-porqu.html	36
7	http://devaneiosdesintericos.blogspot.com/2007/04/polnia.html	37
8	http://fabulas1.blogspot.com/2004/08/29-regras-para-bem-escrever-portugus.html	29
9	http://encarnados.blogspot.com/2007/07/crnicas-encarnadas-200708.html	57
10	http://eroticidades.blogspot.com/2005/06/carta-de-adeus.html	26
11	http://www.abrupto.blogspot.com/index.html	23
12	http://origemdasespecies.blogspot.com/2006/09/prs.html	22
13	http://meianoitetododia.blogspot.com/2007/08/um-corrente-iniciada-por-mim-sempre.html	23
14	http://cidadesurpreendente.blogspot.com/2006/06/aliados-memria-presente_21.html	22
15	http://o-espectro.blogspot.com/2006/03/uma-santanete.html	21
16	http://bichanosdoporto.blogspot.com/2007/03/colnia-de-nevogilde.html	21
17	http://ablasfemia.blogspot.com/2007/04/dupla-precauo.html	20
18	http://abrupto.blogspot.com/2007/11/estado-do-abrupto-o-abrupto-teve-hoje.html	21
19	http://panterasrosa.blogspot.com/2007/11/panteras-rosa-apoiam-cerveja-tagus.html	20
20	http://abrupto.blogspot.com/2007/12/o-meio-eu-li-o-livro-de-carolina.html	19
21	http://causa-nossa.blogspot.com/2007/10/tratado-5.html	20
22	http://quartarepublica.blogspot.com/2007/10/4r-quarta-repblica-o-livro.html	19
23	http://doportugalprofundo.blogspot.com/2007/06/desta-vez.html	19
24	http://esplanar.blogspot.com/2006/03/margarida-rebelo-pinto.html	18
25	http://odoloeventual.blogspot.com/2006/04/grandes-dramas-judicirios-urbino-de.html	34
26	http://ablasfemia.blogspot.com/2006/09/pedro-arroja-no-blasfmias.html	18
27	http://oficiodiario.blogspot.com/search?q=sobre+a+poesia	18
28	http://causa-nossa.blogspot.com/2007/09/quando-se-esquecem-os-princpios.html	19
29	http://sorumbatico.blogspot.com/2007/11/eles-esto-doidos.html	20
30	http://osdiasuteis.blogspot.com/2007/09/futuro-mais-que-perfeito.html	17
31	http://portugalcontemporaneo.blogspot.com/2007/08/lobby-gay.html	17
32	http://ocanhoto.blogspot.com/2007/05/incomoda-me.html	16
33	http://gloriafacil.blogspot.com/2007/01/das-boas-pessoas.html	16
34	http://portugalcontemporaneo.blogspot.com/2007/05/na-pista-do-liberalismo.html	16
35	http://bomba-inteligente.blogspot.com/index.html	16
36	http://doportugalprofundo.blogspot.com/2006/01/perseguio-poltica-estado-de-direito-e.html	16
37	http://ocanhoto.blogspot.com/2007/05/morder-o-isco.html	15
38	http://avenida-dos-aliados-porto.blogspot.com/2005/09/o-parecer-do-ippar.html	15
39	http://ablasfemia.blogspot.com/2007/04/cravos.html	15
40	http://daliteratura.blogspot.com/2007/09/os-que-no-mudaram.html	16
41	http://portugalcontemporaneo.blogspot.com/2007/10/na-massa-do-sangue.html	15
42	http://soc-anonima.blogspot.com/2006/04/os-blogues-de-engate.html	15
43	http://tugir.blogspot.com/2006/09/1.03.html	17
44	http://causa-nossa.blogspot.com/2007/10/referendo-5.html	16
45	http://omundoperfeito.blogspot.com/2007/02/geometria-do-desejo-masculino.html	14
46	http://divasecontrabaixos.blogspot.com/2006/03/v-edio-do-concurso-o-escritor-famoso.html	16
47	http://ablasfemia.blogspot.com/2006/01/referendo-ota-e-tgv.html	13
48	http://geracao-rasca.blogspot.com/2006/12/melhor-blog-temtico-2006top-10.html	13
49	http://ablasfemia.blogspot.com/2007/03/interveno-e-primeiras-reaces.html	13
50	http://ablasfemia.blogspot.com/2006/10/leo-strauss.html	13

Tabela A.4: Ranking of Blog Entries by Incoming Links

A.5 Websites by References

ranking	url	references
1	photos1.blogger.com	587494
2	fotos.sapo.pt	199268
3	bp3.blogger.com	195925
4	bp1.blogger.com	195592
5	bp2.blogger.com	195372
6	bp0.blogger.com	194779
7	en.wikipedia.org	115852
8	pt.wikipedia.org	86632
9	flickr.com	59020
10	youtube.com	41866
11	mailto	38171
12	technorati.com	29174
13	photobucket.com	27543
14	picasa.google.com	23742
15	imdb.com	21223
16	dn.sapo.pt	18545
17	blogger.com	17222
18	myspace.com	15514
19	rapidshare.com	13627
20	jn.sapo.pt	12665
21	javascript	12239
22	imageshack.us	10934
23	diariodigital.sapo.pt	10918
24	publico.clix.pt	10799
25	feeds.feedburner.com	10616
26	dailymotion.com	8578
27	buzznet-00.vo.llnwd.net	7444
28	slide.com	7302
29	correiomanha.pt	6399
30	zerozero.pt	6309
31	portugaldiario.iol.pt	6194
32	videos.sapo.pt	5959
33	the.taoofmac.com	5840
34	ultimahora.publico.clix.pt	5794
35	images.google.pt	5635
36	mediafire.com	5422
37	record.pt	5059
38	imagefap.com	5014
39	blogs.sapo.pt	4568
40	del.icio.us	4533
41	sic.sapo.pt	4462
42	dgsi.pt	4386
43	dre.pt	435
44	news.bbc.co.uk	4344
45	sol.sapo.pt	4311
46	correiodamanha.pt	4208
47	ablasfemia.blogspot.com	4215
48	rapidshare.de	4083
49	www1.folha.uol.com.br	4020
50	tags.sapo.pt	3983

Tabela A.5: Ranking of Referenced Websites

Apêndice B

The Portuguese Blogspace in 2007

Based on Google's Zeitgeist[1], we try to compile our results into some simple data that can be seen as the highlights of the portuguese blogspace over the last year. From our manual inspection of the news archive of the corresponding months, only June presents spam in the results, with a wide set of terms coming from spam blogs that were not filtered by our rules for filtering spam terms.

[1] http://www.google.com/press/zeitgeist.html

B.1 January 2007

ranking	topic
1	saddam hussein enforcamento
2	fiama hasse brandão
3	dakar etapa
4	excomunhão tarcísio cónego
5	biológico adoptivo

Tabela B.1: Topics for January 2007

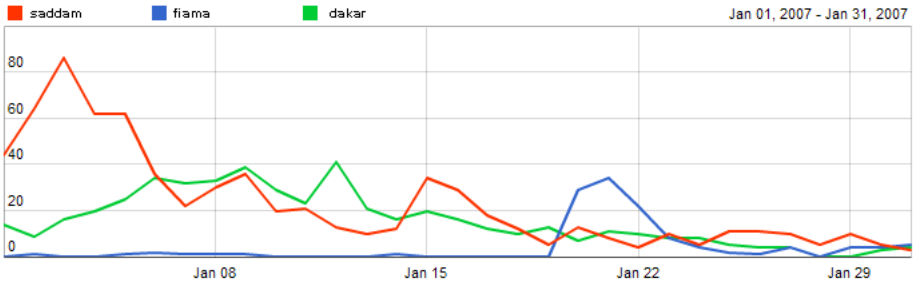

Figura B.1: Results for “saddam”, “fiama” and “dakar” for January 2007

natal	frio	saddam	atlético
fiama	cicarelli	hussein	iphone
hasse	dakar	abortar	marcelo
feto	derlei	excomunhão	tarcísio
apple	biológico	adoptivo	brandão
etapa	champanhe	reveillon	passas
cónego	árbitro	sargento	bessa
resoluções	davos	enforcamento	dragão
artifício	macedo	carmona	sequestro
maternidade	nevar	eliminatória	hillary

Tabela B.2: 40 Most Relevant Terms for January 2007

B.2 February 2007

ranking	topic
1	referendo aborto abstenção despenalização vinculativo eleitores
2	referendo aborto feto
3	referendo votantes
4	referendo aborto votar voto despenalização abortos abortar clandestino
5	zeca afonso
6	scorsese mirren departed whitaker helen oscar martin dreamgirls hudson babel

Tabela B.3: Topics for February 2007

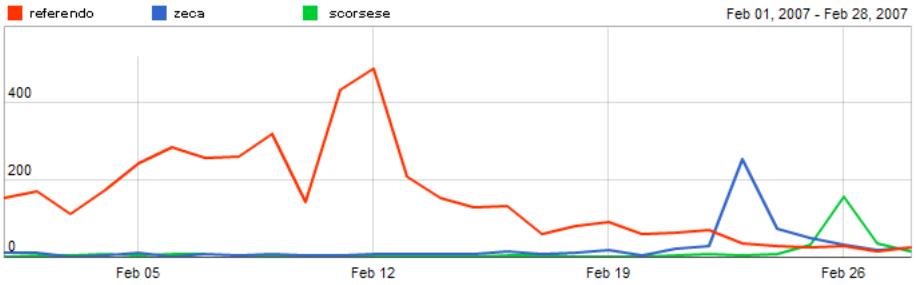

Figura B.2: Results for “referendo”, “zeca” and “scorsese” for February 2007

referendo	carnaval	aborto	namorados
votar	abstenção	voto	valentim
zeca	despenalização	afonso	scorsese
óscares	pinho	abortos	mirren
departed	abortar	clandestino	demissão
vinculativo	salários	chelsea	whitaker
óscar	feto	votantes	arkin
eleitores	helen	oscar	martin
sismo	cinzas	valentine	desfile
urnas	dreamgirls	hudson	babel

Tabela B.4: 40 Most Relevant Terms for February 2007

B.3 March 2007

ranking	topic
1	mulher mulheres internacional 1857
2	bento manuel galrinho
3	salazar museu comba
4	salazar cunhal
5	eclipse lua

Tabela B.5: Topics for March 2007

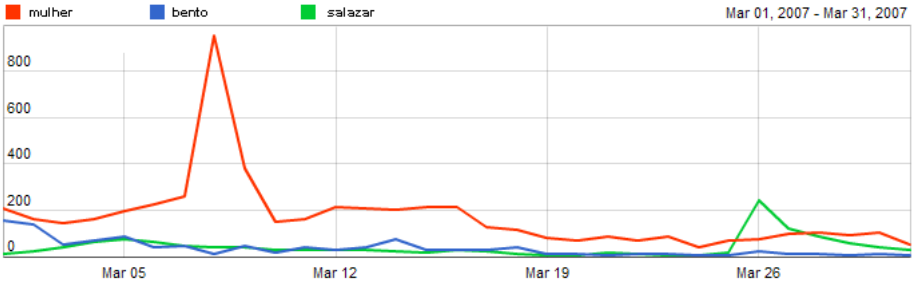

Figura B.3: Results for “mulher”, “bento” and “salazar” for March 2007

mulher	mulheres	internacional	bento
fevereiro	manuel	salazar	eclipse
opa	portas	direitos	bush
rtp	redes	pt	regresso
lua	assembleia	americano	cunhal
1857	braga	galrinho	cds
igualdade	museu	george	accionistas
consumidor	académica	rodrigues	liedson
carnaval	minuto	feminino	baliza
lunar	odete	pp	comba

Tabela B.6: 40 Most Relevant Terms for March 2007

B.4 April 2007

ranking	topic
1	25 grândola morena
2	25 liberdade revolução cravos
3	arguido bragaparques
4	sarkozy ségolène royal

Tabela B.7: Topics for April 2007

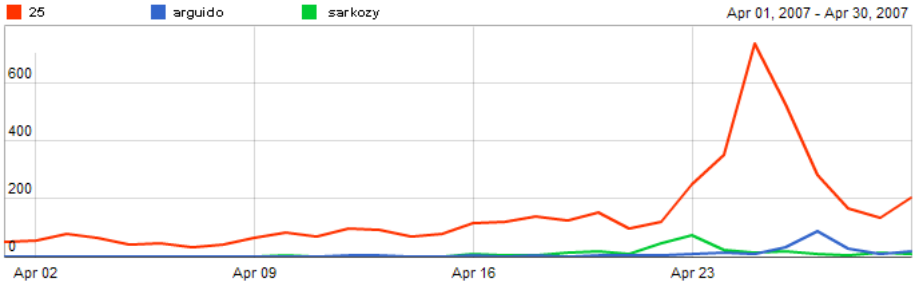

Figura B.4: Results for “25”, “arguido” and “sarkozy” for April 2007

25	liberdade	revolução	páscoa
1974	cravos	carmona	33
feriado	túnel	arguido	cravo
bragaparques	sarkozy	manifestação	comemorações
zeca	ditadura	ségolène	marquês
derby	manifestantes	fascistas	presidenciais
carmo	boavista	cavaco	guernica
grândola	pen	morena	pide
psp	eusébio	fascismo	royal
rangel	armadas	jamor	bayrou

Tabela B.8: 40 Most Relevant Terms for April 2007

B.5 May 2007

ranking	topic
1	dren suspenso directora licenciatura insulto apelida
2	sarkozy royal ségolène franceses
3	rctv venezuela
4	madeleine 282

Tabela B.9: Topics for May 2007

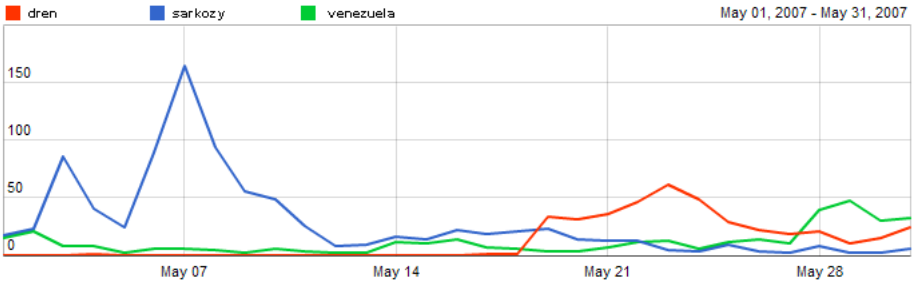

Figura B.5: Results for “dren”, “sarkozy” and “venezuela” for May 2007

greve	lino	dren	sarkozy
carmona	trabalhador	campeão	suspenso
deserto	margem	feriado	aeroporto
directora	licenciatura	ota	seara
roseta	rctv	disciplinar	hergé
royal	insulto	negrão	charrua
madeleine	jocoso	ségolène	tejo
1886	282	presidenciais	telmo
jamor	fumadores	venezuela	camelos
apelida	suspensão	franceses	calheiros

Tabela B.10: 40 Most Relevant Terms for May 2007

B.6 June 2007

ranking	topic
1	attorney furniture condo ringtones attorneys mattress lawyers lxdirect coat jacket lai leather property rental lawyer phones county boots drawer accident cheap sale skirt injury jackets dresses trousers apartments bedstead beach
2	epal epul

Tabela B.11: Topics for June 2007

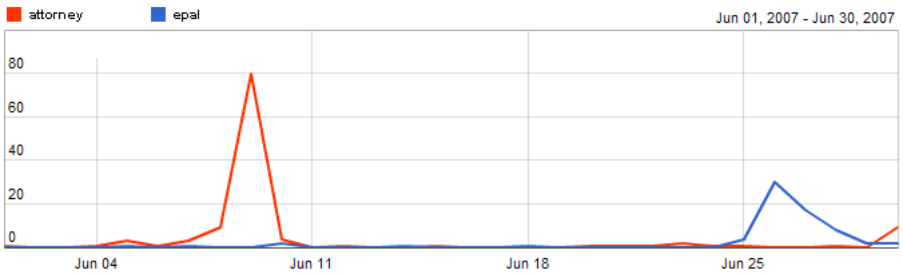

Figura B.6: Results for “attorney” and “epal” for June 2007

berardo	camões	feriado	attorney
furniture	condo	ringtones	attorneys
mattress	lawyers	lxdirect	coat
jacket	lai	leather	ccb
property	rental	lawyer	phones
county	boots	drawer	accident
cheap	sale	skirt	alive
injury	epal	jackets	renan
7.maravilhas	cortesias	dresses	epul
trousers	apartments	bedstead	beach

Tabela B.12: 40 Most Relevant Terms for June 2007

B.7 July 2007

ranking	topic
1	maravilhas redentor cristo petra muralha machu jordânia mahal taj china coliseu picchu estátua méxico
2	abstenção eleições votos
3	eleições lisboetas
4	eleições carmona intercalares
5	eleições carmona roseta votos negrão helena telmo
6	simão sabrosa
7	earth live
8	antonioni michelangelo

Tabela B.13: Topics for July 2007

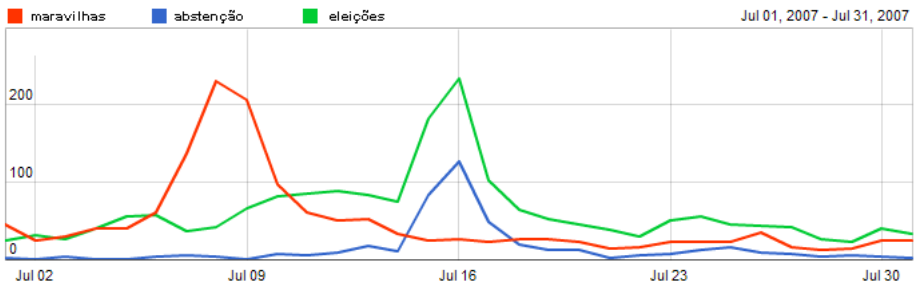

Figura B.7: Results for “maravilhas”, “abstenção” and “eleições” for July 2007

maravilhas	abstenção	eleições	carmona
redentor	simão	earth	roseta
antonioni	live	cristo	lisboetas
petra	votos	muralha	machu
jordânia	mahal	arcade	intercalares
taj	michelangelo	china	maravilha
coliseu	itzá	picchu	eleitores
negrão	estátua	derrota	méxico
helena	telmo	chichén	cabeceiras
bergman	sabrosa	belém	tam

Tabela B.14: 40 Most Relevant Terms for July 2007

B.8 August 2007

ranking	topic
1	prado coelho eduardo epc
2	milho transgénico silves transgénicos eufémia hectare
3	milho transgénico plantação agricultor
4	triplo évora atletismo salto osaka
5	puerta sevilha
6	rodrigues dalila
7	antonioni bergman ingmar

Tabela B.15: Topics for August 2007

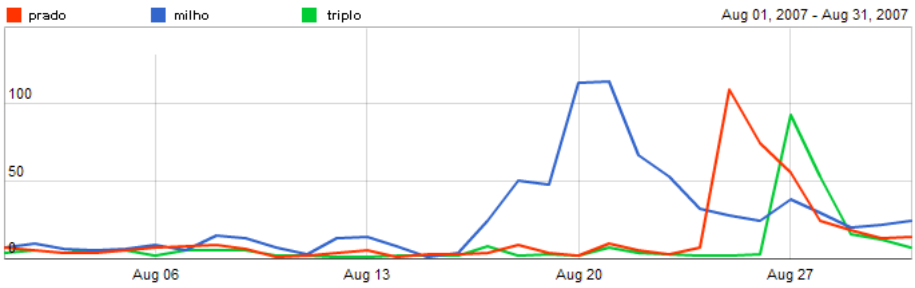

Figura B.8: Results for “prado”, “milho” and “triplo” for August 2007

prado	milho	transgénico	camacho
coelho	eduardo	triplo	évora
silves	feriado	puerta	atletismo
transgénicos	rodrigues	antonioni	bergman
salto	eufémia	osaka	dalila
nelson	gnr	torga	ingmar
indirecto	elvis	tam	sudoeste
nélson	plantação	arménia	sevilha
proença	epc	hectare	árbitro
somague	lance	agricultor	empate

Tabela B.16: 40 Most Relevant Terms for August 2007

B.9 September 2007

ranking	topic
1	scolari soco
2	scolari murro
3	scolari seleccionador fpf
4	scolari seleccionador sérvio sérvia agressão dragutinovic
5	psd militantes quotas
6	psd menezes directas eleições
7	santana lopes chelsea mourinho interrompido
8	chelsea mourinho abramovich
9	mccann maddie madeleine kate gerry
10	aquilino panteão
11	birmània monges myanmar
12	dalai lama
13	pavarotti luciano

Tabela B.17: Topics for September 2007

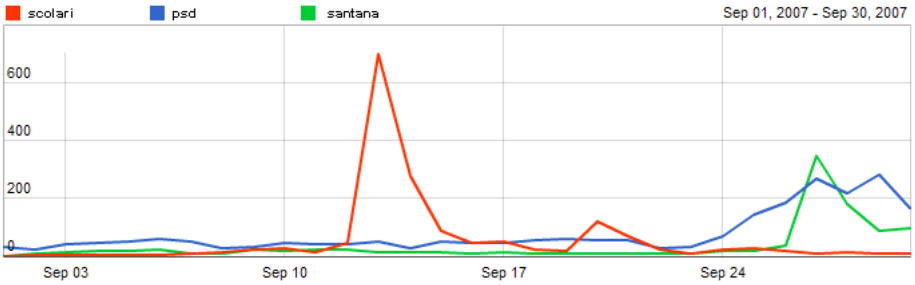

Figura B.9: Results for “scolari”, “psd” and “santana” for September 2007

scolari	psd	santana	menezes
lopes	chelsea	seleccionador	sérvio
directas	eleições	mourinho	mccann
maddie	sérvia	aquilino	birmânia
dalai	militantes	fátima	madeleine
lama	quotas	murro	interrompido
panteão	pavarotti	ahmadinejad	luciano
avante	monges	kate	agressão
abramovich	myanmar	fpf	quaresma
soco	gerry	dragutinovic	escócia

Tabela B.18: 40 Most Relevant Terms for September 2007

B.10 October 2007

ranking	topic
1	tratado referendo reformador
2	nobel lessing doris
3	che guevara
4	pobreza erradicação
5	sindicato covilhã sprc psp
6	sputnik satélite
7	burma birmània
8	habacuc guillermo honduras
9	procurador pgr

Tabela B.19: Topics for October 2007

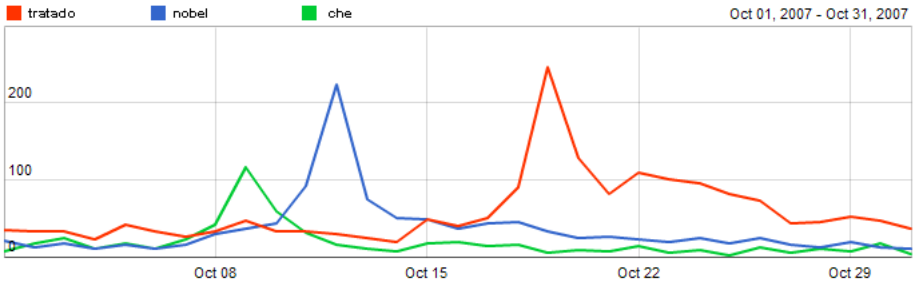

Figura B.10: Results for “tratado”, “nobel” and “che” for October 2007

tratado	congresso	nobel	che
guevara	pobreza	sindicato	referendo
putin	bpi	sputnik	covilhã
adriano	lessing	doris	burma
porreiro	greve	1910	habacuc
feriado	reformador	birmânia	guillermo
cazaquistão	ucranianos	sprc	kiev
psp	procurador	uefa	halloween
action	honduras	pgr	watson
erradicação	monarquia	satélite	directas

Tabela B.20: 40 Most Relevant Terms for October 2007

B.11 November 2007

ranking	topic
1	scolari finlàndia euro apuramento sérvia seleccionador polónia
2	greve valorsul
3	martinho magusto
4	apuramento selecções
5	cozinheiro portador
6	cozinheiro hiv
7	ibero zapatero
8	ctt fónix
9	norman mailer
10	pnud idh

Tabela B.21: Topics for November 2007

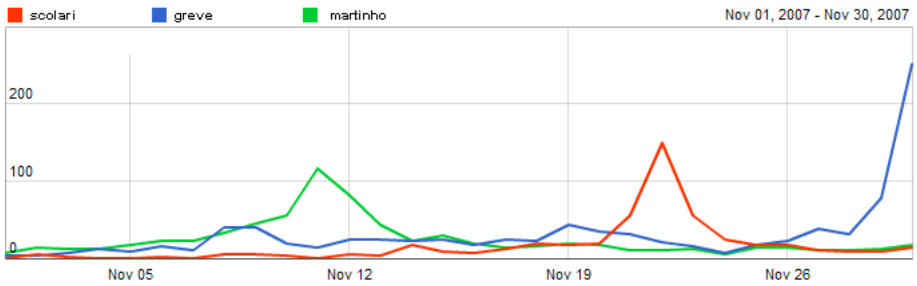

Figura B.11: Results for “scolari”, “greve” and “martinho” for November 2007

scolari	greve	halloween	martinho
finlândia	euro	apuramento	finados
feriado	milan	cozinheiro	sérvia
ibero	bruxas	liverpool	seleccionador
hiv	mesquita	ctt	manchester
uefa	esmeralda	thanksgiving	magusto
norman	defuntos	bernardino	suiça
chover	polónia	fónix	valorsul
annapolis	176	pnud	idh
zapatero	mailer	selecções	portador

Tabela B.22: 40 Most Relevant Terms for November 2007

B.12 December 2007

ranking	topic
1	natal feliz festas santo deseja próspero
2	natal feliz desejar quadra
3	natal feliz christmas merry
4	natal feliz presépio
5	cimeira tratado áfrica ue
6	cimeira áfrica africanos mugabe ue ditadores darfur kadhafi
7	benazir bhutto paquistão musharraf
8	tratado venezuela chávez hugo referendo
9	tratado ue assinatura reformador referendo
10	venezuela chávez hugo chavez derrota venezuelano venezuelanos referendo
11	bcp cgd

Tabela B.23: Topics for December 2007

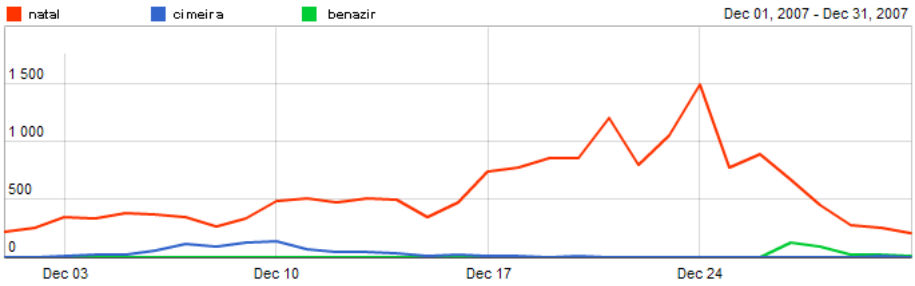

Figura B.12: Results for “natal”, “cimeira” and “benazir” for December 2007

natal	feliz	cimeira	benazir
tratado	festas	bhutto	áfrica
novembro	venezuela	chávez	santo
desejar	paquistão	africanos	mugabe
bcp	deseja	ue	christmas
assinatura	hugo	chavez	merry
derrota	cgd	quadra	próspero
niemeyer	venezuelano	taxa	reformador
ditadores	venezuelanos	musharraf	darfur
presépio	cpmf	referendo	kadhafi

Tabela B.24: 40 Most Relevant Terms for December 2007