International Journal of Electronic Commerce Studies
Vol.3, No.2, pp.335-344, 2012
doi: 10.7903/ijecs.1100

YOUTUBE.PT: A PORTUGUESE PROFILE ON YOUTUBE

João Luís Canais
ISEGI, Universidade Nova de Lisboa
1070-312 Lisboa - Portugal
joao@canais.com

Miguel Neto
ISEGI, Universidade Nova de Lisboa
1070-312 Lisboa - Portugal
mneto@isegi.unl.pt

ABSTRACT

Over the last decade and the spread of broadband network access, the Internet has become the dominant means of distributing multimedia content of excellence. In particular, the emergence of online video publishing and sharing services is now one of the centers of attention on the Internet and allows users to share their content with large audiences.

Available since February 2005, YouTube is the largest online video community with more than 2 billion page views per day. However, a web-based application is, by definition, above any culture, geography, or ideology, treating all equal without any type of distinction.

With this research we attempted to determined the Portuguese profile of YouTube users. Is it possible to discover a pattern of Portuguese content on YouTube? What is the profile of these people? To make this possible we searched for Portuguese YouTube content through the technological facilities that the platform provides and cataloged these findings in terms of content type and user profile. From this set of information, we extracted quantitative and qualitative information that, after being properly treated and analyzed, enabled us to obtain the information that we sought.

Keyword: Portugal, YouTube, User Profile, User Behavior, Internet Culture, Social Network, Video Publishing and Sharing

1. INTRODUCTION

The present work aimed to create a profile of users and Portuguese content on YouTube. In particular, the research questions included: what are the major content areas or subjects? Is there some geographical dispersion of this content? Is there a typical user profile? A final aim was to define a generic model for characterization a Portuguese profile on YouTube.

This characterization had two major reasons: the first was a sociological based reason; specifically, these researchers were interested in reveal geographic and cultural differences in a social network like YouTube and, second was a technological and business related reason; specifically, to better understand why this type of user behavior allows telecom operators to improve services by optimizing the network traffic that is generated by video servers and by promoting business awareness around the video content distribution[1].

To achieve this goal we crawled the YouTube database using the public YouTube API[2] and developed a custom made software program to perform this task[1, 3, 4].

After this phase, the next step was dedicated to data processing and analysis, as well as identifying patterns and behaviors from the raw dataset. To do this, qualitative and quantitative information was extracted and treated to cover all the aspects of this research.

Finally, we will present the first results from this research, draw some conclusions, and define future analysis on the collected data.

2. CRAWLING YOUTUBE

YouTube stores a large amount of videos every day and the question is how can we find Portuguese-specific content? Our approach to retrieve data on Portuguese videos and authors from YouTube consisted of developing a custom made crawl application[1, 3, 4] that, by using the standard and public YouTube API[2], would retrieve the information that we needed.

The YouTube API is a public access API that allows a client program to perform several operations on the YouTube database and website. It is possible to search for videos, retrieve standard feeds, and see related content. The client program can also act like a user to upload videos, modify user playlists, and more.

Trying to find Portuguese content and authors on YouTube from scratch, using only the search engine, is a risk because of the data quality and ambiguity factor that is present in search engine responses produced from text metadata information. To overcome this problem, our solution was to start crawling YouTube and use each individual spatial coordinates' data to test against the official Portuguese postal codes geographic coordinates list.

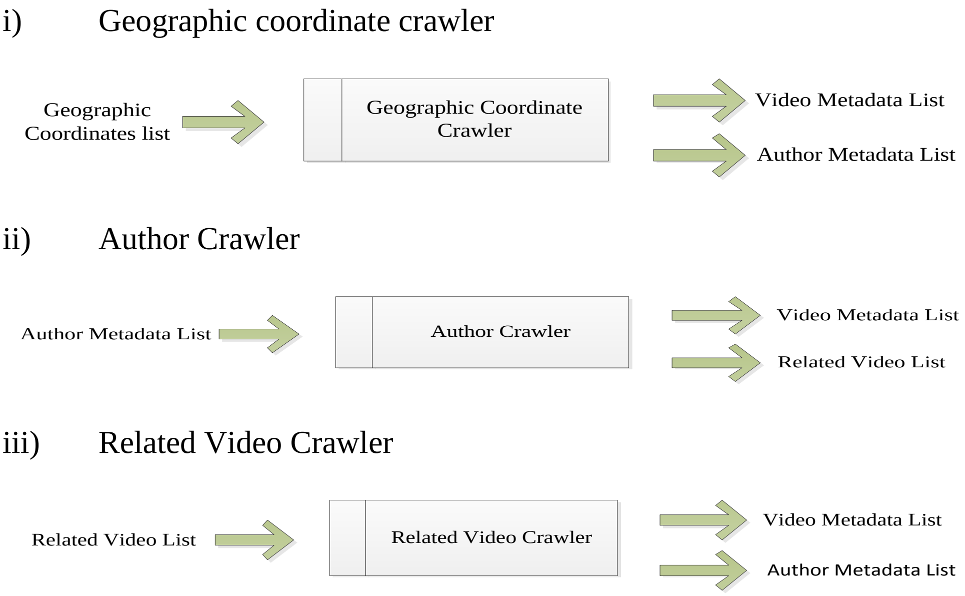
The official Portuguese postal code list identifies each street or district in a city univocally, which allows for the fast and secure delivery of postal mail to each address. This basic concept is the ideal starting point to creating a 'starting set' of videos that, assuming some margin of error, is true Portuguese content. Using this list, each postal code was translated into a unique geographic coordinate (using a public Internet web service) that theoretically identified the geographic center of the street or district. Using this list as our 'starting set,' our custom crawler engine queried the YouTube database to each individual coordinate in search for geographic-specific videos.

Each YouTube video provides a number of different sources to discover other YouTube videos such as the user video list, favorite video list, and related videos (videos identified by YouTube algorithm), and so on. Starting from the video 'starting set' (postal code generated video list) we chose to use as data sources the user video list and related video list to crawl in several cycles going deeper in the tree of relations for this network.

For each video crawled, we collected all the information possible: video id, author id, title, description, keywords, media file information, categories list, ratings, view counts, and comments for each video[5]. Each new video was evaluated to determine whether it was a Portuguese video or not. This evaluation was via several parameters (title and description contained some Portuguese keywords, the video contained Portuguese geographic coordinates or the video author was identified as a Portuguese citizen) that were then combined as a numeric factor of trust for the video content.

For each individual author (or user) crawled, we collected the following information: name, age, country, personal information and interests, creation and last visit date, video count, view count, and subscriber count. For each new author found, its verified if their country value was set to Portugal. In this case, the author was selected for a full crawl of the user video list and related video list, if not, the author was inserted in the database for account proposes.

From a functional point of view, our crawl system was composed of three separated process that shared the same information sources:


i) Geographic coordinate crawler

ii) Author Crawler

iii) Related Video Crawler

Figure 2. Crawling strategies (high-level overview)

We started crawling YouTube on February 11, 2011 and estimated to stop crawling by the end of March. After 45 days of crawling, our database was 16 GB size and the number of crawled information is described in Table 1.

Table 1. Crawled items

	All Crawled	Evaluated as Portuguese Content
Videos	770083	481906 (62%)
Authors	287282	49610 (17%)

3. DATA PROCESSING AND ANALYSIS

For each crawl cycle, thousands of videos and authors references (or metadata) were stored in our database. All information collected by the crawl engine could be characterized in two types of information: qualitative and quantitative information[6].

Quantitative information or ‘factual data,' such as the number of views or video duration were objective facts about something. The qualitative information included titles, descriptions, tags and keywords, were, by definition less trustworthy for this type of study, despite the characteristic that our data analysis needed to cover both quantitative and qualitative information to cover all aspects of the research.

The first step of data analysis, which is presented in the present work will be to define several types of tops lists and linear graphics to identify the basic profile of our sample. The next step in our data analysis will be to identify patterns and behaviors among the data; this type of analysis will provide us the tools necessary to define a model of social network relationship.

A major concern in our analysis process is the data quality analysis. This analysis is required because our crawl process introduced some error in the several decisions made during the crawl phase (see diagrams in annex). Another factor is the data quality itself, since our sample data was taken from an Internet website it is not granted that this data does not contain several types of data malformations[7] (example: some videos that are not Portuguese content have geographic coordinates inside Portugal).

Finally, the need to remove data ambiguity from the free text fields (example: unify 'Lisboa' and 'Lisbon' or 'Porto' and 'Oporto').

4. RESULTS

In this section, we will present some preliminary results that were derived from several direct queries made to the raw data dataset to give us some insight into the Portuguese profile on YouTube.

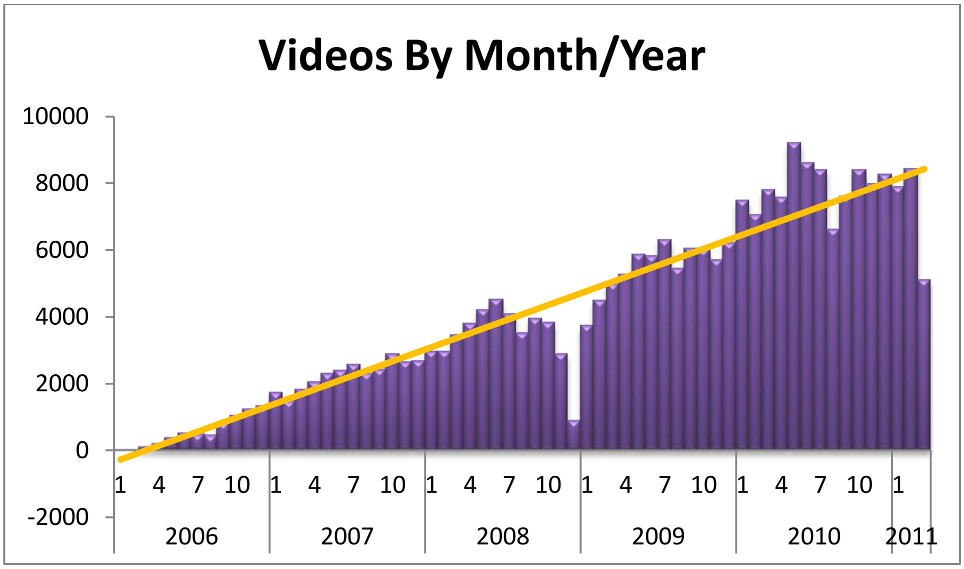


Figure 3. The video creation rate by month and year

Figure 3 show how Portuguese videos increased since 2006 for a total of 256,378 videos by the end of 2010. This rate of increasing was constant, except for during the winter of 2008/2009 (since these results are preliminary at this time we cannot provide a feasible justification for this gap).

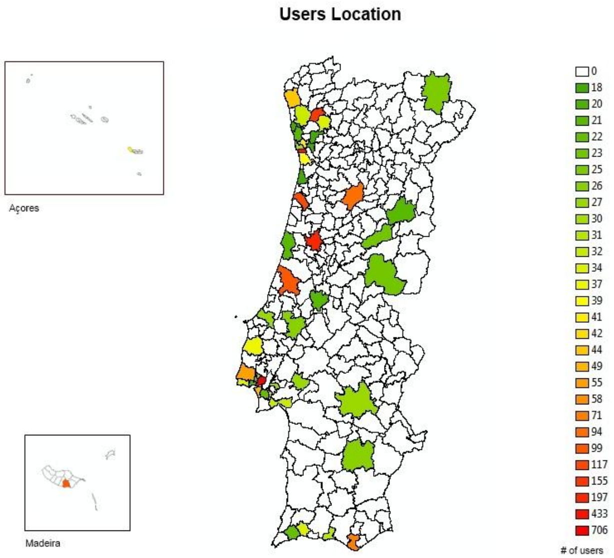


Figure 4. Users location, group by major cities

Figure 4 presents the user (or authors) distribution among the major cities in Portugal. As expected, the major cities or cities near the coast yielded more users than did other cities that are more geographic distant from the coast, which was also consistent with the known statistics on Internet adoption and usage.

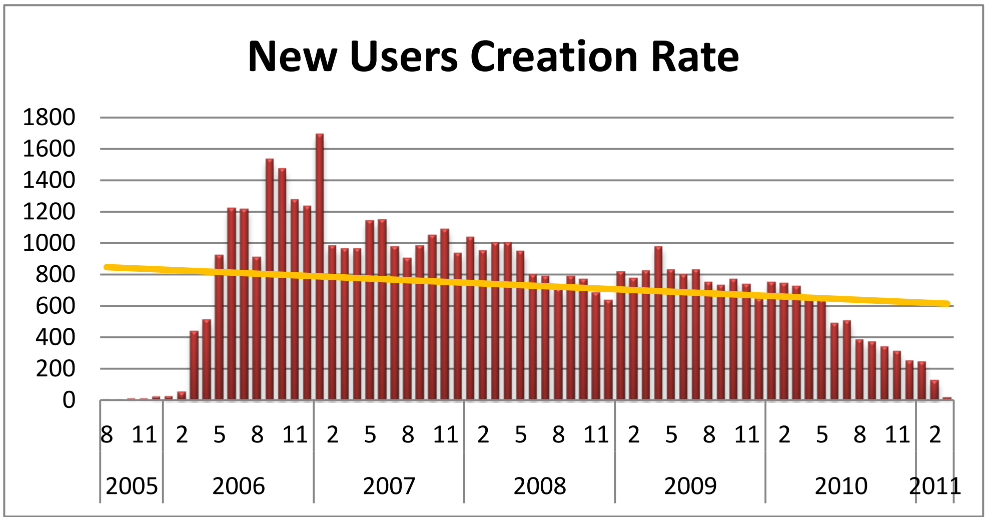


Figure 5. New users (or author's) creation rate

In figure 5 shows the new users (or authors) creation rate. Our first reference to a new user was from August 2005; however, a major increase in new users occurred in 2006. Since this time, new users rate decreased slightly until 2009 and faster during 2010. This evolution could be related with the growing usage of other social services for media publishing, maybe a Facebook, as a side effect.

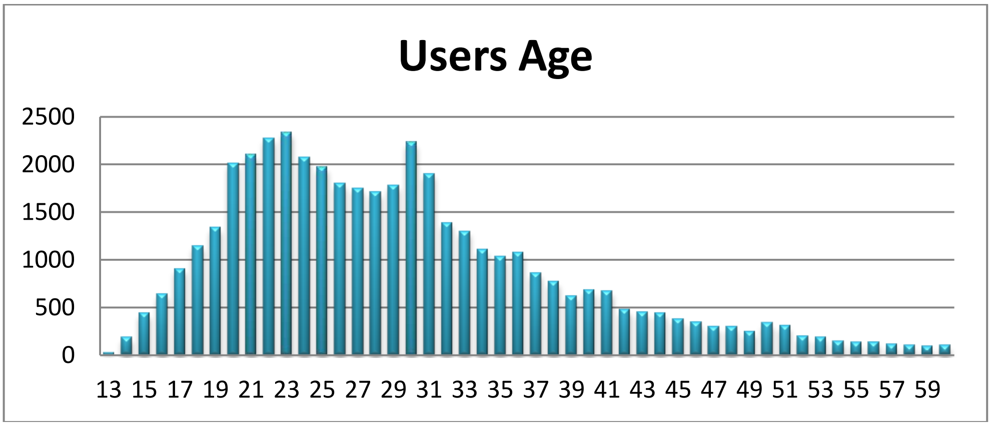


Figure 6. User age distribution

Figure 6 represents the age distribution of our users. As expected the major users groups were from 20 to 32 years old, which follows the known trends of social services usage.

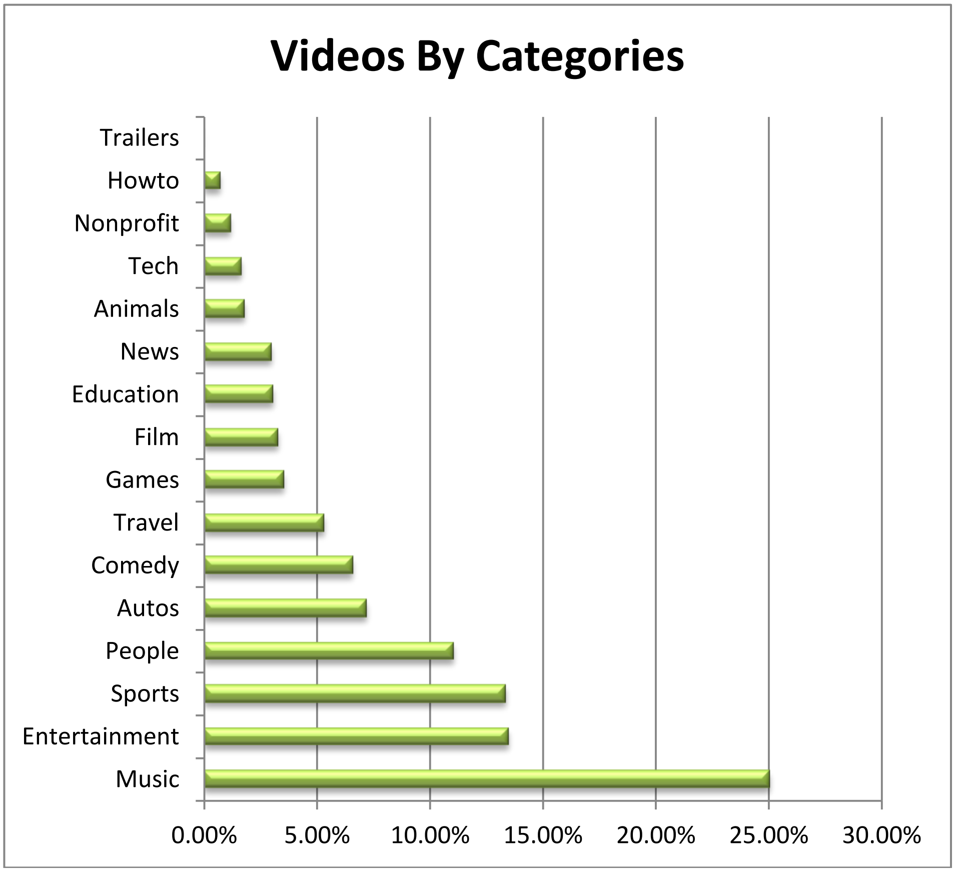


Figure 7. Video categories distribution

Based on the category classifications made by the users in the video upload moment, Figure 7 shows how videos are distributed by YouTube categories. Music was by far the major category; and sports and entertainment had equal values in second place, which shows that users are using YouTube mainly for entertainment and leisure activities.

5. CONCLUSIONS AND FUTURE WORK

During this research, we intended to target two main goals:

1. Understating how YouTube has changed our society globally in the way that content is produced and distributed by each individual.

2. Understanding how YouTube has changed the Portuguese society; measuring the society mobilization for certain types of events, or observing the movement and sharing of information and, above all, trying to find a "soul" in Portuguese YouTube.

To date, we have crawled, successfully, almost half-a-million Portuguese videos and almost 50k Portuguese YouTube authors, these values are good indicators that our crawler strategy was well defined.

The preliminary results derived from several direct queries made to the raw data dataset conform to what is to be expected in this type of Internet service. An example of this is the age distribution and new users creation rate. Additionally, the location distribution is equivalent to the demographic Portuguese distribution (more people live in major cities and near the coast).

In the present step of this research we are moving from the crawling phase to the data analysis phase. Our work is now focusing in validating and analyzing the sample data and defining the best strategies for extracting relevant information from this raw dataset.

As referred, the next step in our data analysis will be identifying patterns and behaviors among our data to provide us better knowledge of social network relationships of this media.

6. REFERENCES

[1] F. Duarte, F. Benevenuto, V. Almeida, and J. Almeida, Geographical characterization of YouTube: A Latin American view. In Virgílio A. F. Almeida and Ricardo A. Baeza-Yates (Eds.), *Proceedings of Fifth Latin American Web Congress* (p13-21). Santiago de Chile: IEEE Computer Society, 2007. doi:10.1109/LA-Web.2007.17.

[2] Google Code, YouTube APIs and Tools. Google.com. Retrieved March 20, 2011, from http://code.google.com/intl/pt/apis/youtube/overview.html.

[3] D. Chau, S. Pandit, S. Wang, and C. Faloutsos, Parallel crawling for online social networks. *Paper Presented at the 16th international conference on World Wide Web*, Banff, Alberta, Canada, May 8-12, 2007. doi:10.1145/1242572.1242809.

[4] K. Lai, and D. Wang, A measurement study of external links of YouTube. *Paper Presented at* the 28th IEEE conference on Global telecommunications, Honolulu, Hawaii, USA, November 30-December 4, 2009. doi:10.1109/GLOCOM.2009.5426136.

[5] X. Cheng, C. Dale, and J. Liu, Statistics and social network of YouTube videos. *Paper Presented at the 16th International Workshop on Quality of Service*, Enschede, The Netherlands, June 2-4, 2008. doi:10.1109/IWQOS.2008.32.

[6] K. Haamer, Africa on YouTube: Social Media's Global View of the PALOP Countries, Tallinn University - Baltic Film and Media School, Tallinn, Estónia, 2010.

[7] E. Felinto, VIDEOTRASH: O YouTube a Cultura do “Spoof” na Internet. In: XVI Encontro da Compós, Junho 2007, UTP, em Curitiba, PR, Brasil.