TWITTERECHO: DISTRIBUTED *CRAWLER* FOCUSED ON PORTUGUESES AT TWITTER

Eduardo Jorge Silva Leite de Oliveira

Project/Dissertation developed under supervising of Eduarda Mendes Rodrigues and Luís Sarmento

Faculdade de Engenharia da Universidade do Porto

1. Motivation

Web has brought applications which allow users to communicate with each other. It is well shown for example in social networks like Facebook, Twitter and others. The growing popularity of social networks transforms them in an excellent way of marketing, information diffusing, studies of opinion and others. Collecting data from social networks is getting more and more important due to run different kinds of studies on them.

2. Problem description

The goal of this dissertation is to collect data from Portuguese Twitter users. It is a complex problem. In fact, it is a set of distinct and complex problems. The main challenges of the created system are:

- Twitter restrictions: maximum number of calls to get data.
- Continues data collecting: the system has to be able to keep working for a long time (at least couple months).
- Easiness of updating: due to the need of continues data collecting, is not acceptable to turn off the system in order to run updates.
- Identifing target users: identifing a Portuguese user is not a trivial task. For example some Portuguese's write in English and Portuguese language is also used in other countries (example Brazil).
- Data size: is intended that beyond actual state of users, all evolution of relations, tweets and stats are store.
- Reliability: errors in collected data will affect studies on it.

3. TwitterEcho

The TwitterEcho tries to solve the previously presented problems. TwitterEcho is a distributed system. The server stores data and clients interact with it and Twitter, as it shown on figure 1.

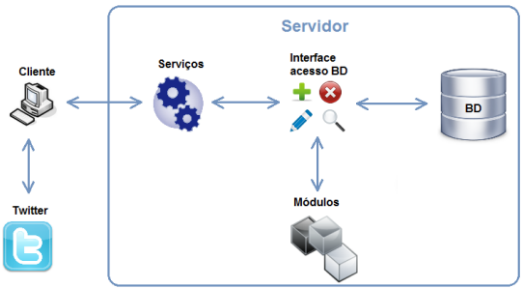

Fig. 1 –TwitterEcho arquitecture.

3.1. Clients

One client communicates with server through services, sever make operation in DB.

The system is scalable to support multiple clients (functions and quantity).

Actually system has two clients with distinct functions: collecting tweets and collect relations (friends and followers).

3.2. Database

TwitterEcho stores large volume of data (GB's of data and millions registers). To allow search and statistics treatment, the data is stored in a relational database management system MySQL.

The server keeps information sent by clients: users profiles, tweets, statistics, list of friends and followers. In order to detect faults the server saves: errors caused by itself or clients, last Access of each client and evolution of DB tables size. It also stores information to grow the users and select them.

3.3. Services

Services are responsible to coordinate clients and get data from them. There are following services used in TwitterEcho: *get*, *put* and *put_error*. *Get* provides list of users which server wants data, *put* receives data and *put_error* receives errors. Communication between system agents are representedon figure 2.

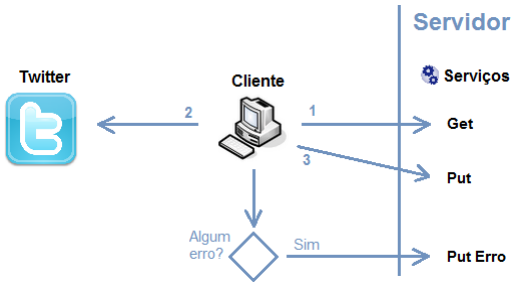

Fig. 2 – Communication between system agents.

Get service uses a scaling algorithm to provide users list. P*ut* service receive data from clients and insert them into DB.

3.4. Modules

Modules are responsible for different tasks. These tasks can be divided into:

- Users grow up: add followers and users mentioned in tweets.
- Users selection: profile analyze and language identification to recognize Portuguese's users.
- User's verification: verify if user's accounts are still valid (not deleted or suspended) and "freeze" of inactive users.
- Statistics creation: annotations of DB grow up.

3.5. Growing

TwitterEcho started working at 27th April 2011 14h40. In the beginning DB had 2052 users at . In 24h 4017 users were automatically added. This value represents approximately 200% grow in 24h.

From 27th April to 17th June, 2011, TwitterEcho collected about 65 thousand users, 2.8 million tweets, 215 thousand follower's lists and 130 thousand friend's lists..

4. Conclusions

TwitterEcho allows to obtain large data from Portuguese Twitter community. It has been achieved by a distributed system, which can also be adaptable to others communities. The implementation of the distributed system allows collecting data beyond the point granted by Twitter.

TwitterEcho is modular, and allows flexible allowing incremental addition of news functionalities. It is robust and reliable, had been working for 45 days. It collect data mostly from Portuguese users, being completely autonomous growth of the number of users to obtain information. It showed good results in evaluation, language identification of Portuguese from Portugal excellent precision (97.4%) and recall of 65.9%. From a large sample of users classified as Portuguese 90.8% of them was correctly classified.

Concluding, in my opinion the objectives of the dissertation have been achieved. A complex and innovative system which is able to make a valuable contribution to depth studies on the Portuguese community on Twitter have been created. Due to the interest in this system, there already have been developed several applications based on data collected by TwitterEcho.