Graph Visualization Tool for Twittersphere users based on a high-scalable Extract, Transform and Load system

Pablo Aragón, Íñigo García and Antonio García

Cierzo Development

pablo@cierzo-development.com
igmorte@cierzo-development.com
antonio@cierzo-development.com

Abstract

Microblogging is a service emerged from Web 2.0 that allows to publish posts of limited length, generally less than 200 characters in a user profile Web page. Most of these services, like Twitter[1], Jaiku[2] or Tumblr[3] provide the option to subscribe to other users, thereby, each post is sent immediately to those who have chosen the option of receiving them. Twitter is the most widely used microblogging system. In it, users publish updates and interact with other members by forwarding updates or quoting users in their posts by mentions or replies.

The last feature makes up a large repository of Web information that includes connections between users by associating each link with a textual content. The application presented in this paper shows how to mine data through a graphical tool that identifies key influencers related to a search, and it organizes the results into a graph structure. Thereby, the final client manages a mechanism for recognizing key opinion makers in a specific subject by visualizing them as interconnected nodes..

Categories and Subject Descriptors H.3.3 [*Information Systems Applications*]: Information Search and Retrieval - Information Filtering; J.4 [*Social And Behavioral Sciences*]: Sociology

General Terms High-scalable Architecture, Extract-transform-load Processes, Social Media Analysis, Microblogging, Twitter

1. Introduction

The structure of a Twitter profile consists mainly of two frames, a central one that forms a wall with the latest published posts (called tweets) and a second one with the author information as its description, geographic location, or the following users. Each of the tweets of the wall contains a set of metadata: update text, publication date or if it is a retweet.

[1] http://twitter.com/

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

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

WIMS'11 May 25-27, 2011 Sogndal, Norway

Most of the main social networks are built to share content between users who belong to their own network. Each user has the feature to ask other users to join his personal network, while they accept or reject such requests. Nevertheless, Twitter is designed to establish unidirectional connections between users. This means that a user can establish a relationship with another through mechanisms that Twitter website describes as[4]:

- Reply: Update that begins with @username.
- Mention: Update that contains @username in the body of the tweet.
- Retweet: Update that contains the body of the another user tweet by specifying the original author.

The first two express a quote from one user to another through the content of the tweet. In this manner, it sets a semantic relationship between a source user and a number of target users associated with the words that constitute the tweet. The extraction of metadata from a tweet and its mining generate a collection of content, users and their relationships.

1.1 State of the Twittersphere

Twitter has become the leading microblogging service with more than 40 million users, surpassing 100 million tweets[1]. This involves not only the reality of the large volume of information to be processed, it is also a set of data continually expanding and highly influenced by social trends such as gender or geographic location[2]. Also, being a real time system, the value of a mashup designed to exploit Twitter data requires an automatic capacity adjustment process.

1.2 Who are the main influencers on a specific topic?

Nowadays, there are several statistical tools for the Twittersphere. Some of them establish metrics based on data as the number of tweets, the number of followers or ratios as the quotient of following and followers of each profile. Such tools can be considered valid for measuring the relevance of users and distinguishing general main influencers. However, these tools do not establish the importance of users who participate in a particular subject or the ones who interact with a specific user.

On the other hand, there are available search engines that crawl the Twittersphere but the lists of results do not set a value to users from the query response. In a social network like Twitter, that value should be highly dependent on the impact generated by the tweets. This impact is measured by extracting and counting men-

[4] http://support.twitter.com/entries/14023-what-are-replies-and-mentions

tions/replies to identify users and connections to other users in the tweets from the response.

The purpose of the application presented in this document is to offer a tool for monitoring the Twittersphere based on a high-scalable system. By conducting a textual search on the content of the tweets or on the users to analyze, the tool generates a graph of the results from the index. The final graph displays the importance of each resulting Twitter user by associating a specific weight to each node and representing the existing relationships with other users.

1.3 Structure of the paper

This paper is organized as follows. The system architecture, its distributed design and modules, is described in Section 2. The results of two use cases are presented in Section 3 by including global graphs, filtered graphs and tables with detailed information from main nodes. The conclusions are in Section 4.

2. System architecture

This section describes the architecture implemented in the Extract, Transform and Load system by focusing on its highly distributed design and its modules.

2.1 Distributed design

The data volume stored in Twitter indicates the need to define an architecture capable of handling large volumes of information at the lowest possible operating cost. On the other hand, the daily growth of the number of users on Twitter, and the oscillations in the frequency of publication, generates a high degree of uncertainty. Therefore, ensuring the scalability of the system requires an automatic processing capacity balancer mechanism. The resolution of both conditions is carried out by a design implemented on distributed computing infrastructure Hadoop in Amazon Elastic Compute Cloud (called Amazon EC2).

Hadoop is a distributed framework for large-scale processing. Its design enables to spread large workloads across a cluster of machines by implementing MapReduce programming paradigm[3] on a distributed file system based on Google File System[4]. In a Hadoop cluster, data is distributed among all the nodes that comprise it. The distributed file system Hadoop splits large files into pieces that are managed by several cluster nodes. Moreover, each fragment is replicated on multiple nodes. Thereby, a failure in a machine does not affect data availability.

Amazon EC2 infrastructure arranges to configure machine images (called AMI) that correspond to the snapshot of features and stored data from a specific server. Storing the AMI of an initial Hadoop node, the system balancer mechanism calculates automatically the real-time computation demand by adding AMI nodes or removing production ones. The system becomes highly sensitive to workload peaks and, thereby, optimizes the cost of a computer system in the cloud.

2.2 Implemented modules

The Figure 1 represents a system overview detailing the technologies of each module.

2.2.1 Crawling module

The crawling module, based on Nutch[5], is the most resource intensive. Nutch is an open source search engine robot managed by the Apache Software Foundation.

Each crawling process consists of two executions. The first one crawls Twitter profiles whose urls are stored in a database. This process extracts outlinks to new profiles that are inserted into the database by implementing a frontier expansion mechanism. In turn, tweets urls are extracted from the wall of each profile and stored in another data structure.

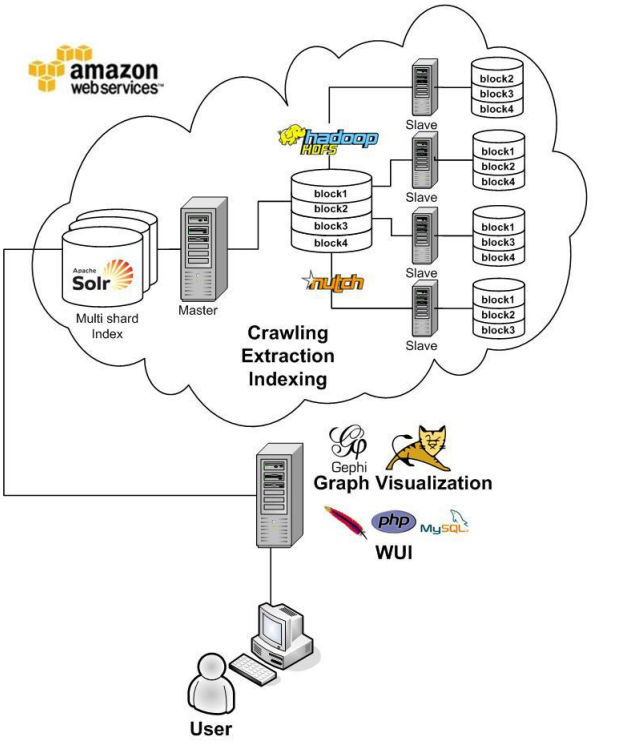

Figure 1. System overview

The frequency of publication between users varies considerably. In a large-scale system as presented here, it is essential to minimize computation costs. Therefore, a ratio is applied to the time that must elapse between successive fetchs to the same profile. This ratio is based on the heuristic that the profile which has been updated has a high potential to add new tweets in a short space of time, while the profiles that have not been updated will have a slightly lower potential in future fetch processes.

In contrast to the blogosphere, where posts are modified by the appearance of comments, a tweet can never be changed after its publication. Therefore, the crawling module performs a single fetch for each tweet. A process calculates the differential of crawled tweets against the ones from past fetchs of each profile. The output of this process contains all new tweets which were collected in the previous crawl. Those ones are processed in a second run to download the HTML content of each update which becomes the input of the metadata extraction module.

2.2.2 Metadata Extraction module

The HTML code from the url of a tweet contains a set of metadata: textual content, publication date, author, and mentions to other users. The metadata extraction module performs lexical analysis through regular expressions on the HTML code of each tweet and obtains such set.

As most global networks, the textual content of tweets can be written in different languages. Operating a monitoring system often requires the implementation of a market segmentation identified with

the idiom. The extraction module, through several languages n-grams profiles, detects the language of each tweet and incorporates it to the metadata set. In addition, tweets keep an upper limit of 140 characters and sometimes languages are mixed in a single update. Thence, the numeric language classifier output is also inserted.

2.2.3 Indexing module

Metadata sets, which are extracted in the previous module, are stored in an index to perform full-text queries that generate the input of the visualization module. The technology used in the indexing module is Apache Solr[5]. Solr is a search engine implemented in Java which wraps Lucene indexing library[6]. This provides an engine system with search results sorting algorithms, stemming, stop words filters or faceted searchs. The module indexes and stores the metadata defined for visualization module searches.

The index architecture is composed by several shards where the appearance of a document in a particular instance is determined by publication date. The monitoring of social media information tends to be at specific intervals of time. Therefore, module indexes and stores close tweets in the same shard performing optimized response times in a large multicore architecture.

2.2.4 Graph Visualization

The Graph Visualization module displays the graphical information through a Web User Interface by Gephi Toolkit[6]. The user enters the type of search he needs: search on textual content or user based search. The module also allows to establish a range of dates of publication of the resulting tweets.

The module generates a graph with the results of search on the index. At that time, the module filters the values from the graph and runs an iterative layout Hu's algorithm[7]. Then, it runs another algorithm to extract the metrics of each node to rank. The ranking node values are essential for calculating color and distance attributes. Finally, the resulting graph is exported to the Web interface.

3. Results

The purpose of this section is to include the tool result for two use-cases.

3.1 Western Sahara Conflict

On November 8, 2010, the Moroccan security forces involved in Izik Gdeim camp, located on the outskirts of the city of Laayoune (Western Sahara). Western Sahara is an African country from United Nations list of Non-Self-Governing Territories whose sovereignty is disputed between the Government of Morocco and Polisario Front. Moroccan action was criticized by various sectors of Spanish society due to its status as colonial territory from 1934 to 1975. The impact on the Spanish-speaking Twittersphere from 8 to 18 December, is analyzed in Figure 2 . It represents the nodes of the 1721 users who published or were mentioned in one of the 3925 tweets that include sahara as a token. The total number of mentions is 707. As the figure shows, there is a high contrast between a large central cluster of interconnected nodes and numerous clusters of smaller nodes around it.

The tool allows to filter the initial graph to represent only nodes with a number of edges greater than or equal to a concrete limit. The

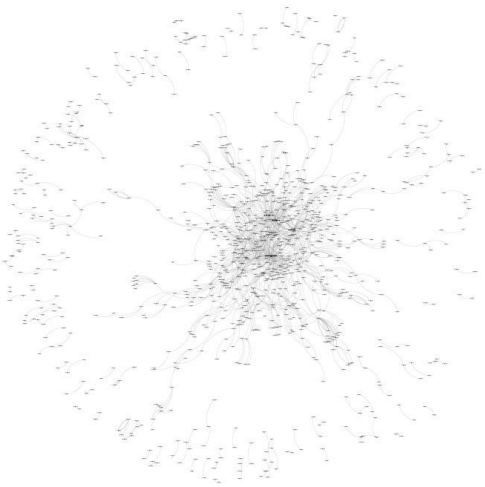

Figure 2. *Western Sahara Conflict* graph

graph presented in Figure 3 is filtered by limit=5. The new graph highlights three nodes which are identified by their Twitter Profile description. They correspond to the users who figure in Table 1 within the number of own tweets, tweets where they are mentioned, and the number of users with whom they are related:

Name	Tweets	Quotes	Users
pepeluibiza	137 (3,49%)	1 (0,14%)	64 (3,72%)
pituskaya	75 (1.91%)	2 (0,28%)	26 (1,51%)
freedomsahara	164 (4,18%)	6 (0,85%)	28 (1,63%)

Table 1. *Western Sahara Conflict* most weighted users

The high relevance of the three users has been corroborated by a qualitative cross-validation performed by social media consultants. PEPELUIBIZA corresponds to an active and anonymous user who publishes political tweets, PITUSKAYA is a Human Rights activist and FREEDOMSHARA corresponds to a Pro-Sahrawi Association.

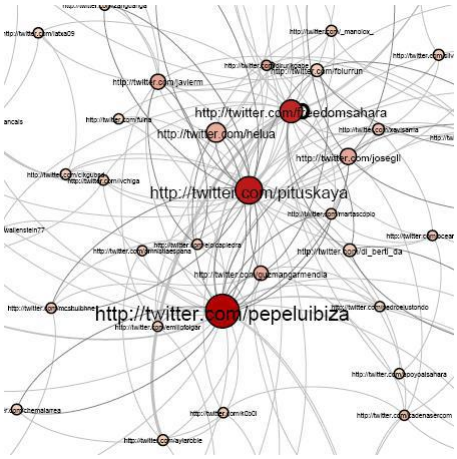

Figure 3. *Western Sahara Conflict* filtered graph

[5] http://wiki.apache.org/solr

[6] http://wiki.gephi.org/index.php/Toolkit_portal

3.2 Patxi López

Patxi López holds the position of President of the Government of the Basque Country (Spanish region). His campaign included strategies in social networks like Facebook[7] or Tuenti[8], and also on Twitter[9]. His political career implies a thought leadership confirmed with more than 90000 followers. Figure 4 represents the graph of a search for tweets that mention Patxi López. The results are a total of 196 tweets including 366 mentions who involved 186 users. The graph nodes correspond to all users that include @patxilopez, either as the author or as a mentioned user in that tweet. The visualization module identifies patxi_lopez as the maximum value node by setting it as the central one.

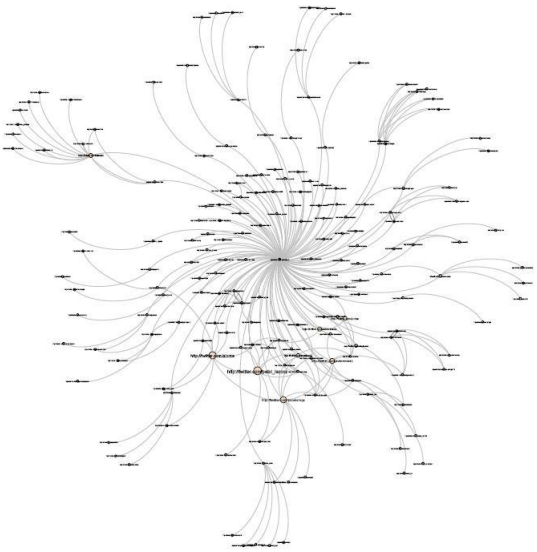

Figure 4. *Patxi López* graph

Again, the graph is filtered for removing nodes with less than 5 edges. Figure 5 presents a reduced set of nodes and Table 2 lists three main nodes data similarly to the previous case:

Name	Tweets	Quotes	Users
patxi_lazcoz	1 (0,51%)	4 (1.09%)	10 (5,37%)
psoelarioja	6 (3,06%)	1 (0,27%)	9 (4,83%)
javierremirez	10 (5,10%)	1 (0,27%)	6 (3,23%)

Table 2. *Patxi López* most weighted users

Also, the relevance of the three users has been corroborated by a qualitative cross-validation performed by social media consultants. PATXI_LAZCOZ is the first Mayor of Vitoria (capital city of the Basque Country) who belongs to the Socialist Party, JAVIERREMIREZ is the New Technologies Socialist Party Secretary, and PSOELARIOJA is the Twitter profile for the Socialist Party for La Rioja (Autonomous Community which borders the Basque Country).

[7] http://www.facebook.com/

[8] http://www.tuenti.com/

[9] http://twitter.com/patxilopez

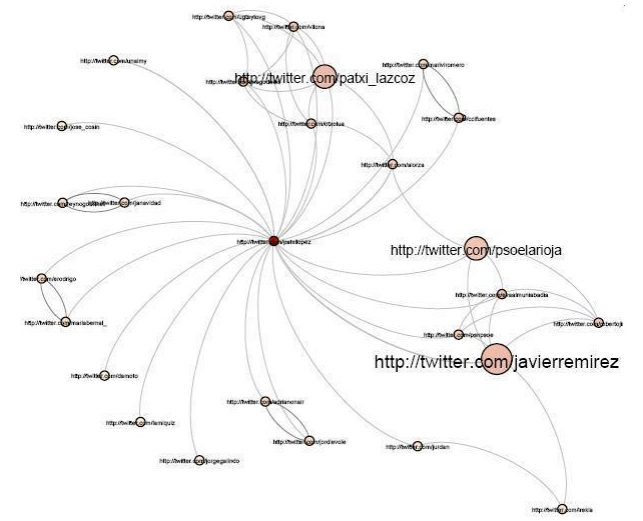

Figure 5. *Patxi López* filtered graph

4. Conclusions

The two use-cases express the simplicity for the end user to identify Twitter users and their connections. The first case reflects the functionality to monitor the users involved in a topic and relations between them. The second use-case displays Twitter users ecosystem around a specific profile.

This paper contains a new approach to visualize and classify Twitter key users in a concrete subject. On the one hand, it provides the integration of crawling, metadata extraction and indexing modules into a system with a high-scalable design for managing large data volumes. On the other hand, it describes the graph visualization tool functionalities for identifying major influencers in a social network.

Future work on this area includes extending the system to other information repositories such as the blogosphere or forums, and developping new data mining applications to increase the extracted knowledge from the Web.

References

[1] M. Cheong and V. Lee, "Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base," 2009.

[2] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" 2010.

[3] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," 2004.

[4] S. Ghemawat, H. Gobioff, and S.-T. Leung, "The google file system," 2003.

[5] R. Khare, D. Cutting, K. Sitaker, and A. Rifkin, "Nutch: A flexible and scalable open-source web search engine," 2004.

[6] O. G. Gospodnetic and E. Hatcher, *Lucene in Action*, Manning, Ed., 2005.

[7] Y. Hu, "Efficient and high quality force-directed graph drawing."